Compilation of Undergraduate Work in Various Subfields of Computer Science and Mathematics

Emerson Kahle

February 10, 2024

Introduction

This is a compilation of various assignments which I (Emerson Kahle) have completed for my undergraduate courses at the University of Southern California. The work includes assignments from

1. *Fundamental Concepts of Analysis*
2. *Numerical Methods*
3. *Mathematical Statistics*
4. *Mathematics of Machine Learning*
5. *Probability Theory*
6. *Algorithms and Computing Theory*
7. *Theory of Numbers*

All of the work (solutions) are entirely written by myself in Overleaf using LaTeX. Assuming that each assignment was thoroughly graded, the work is almost completely accurate.

Note: None of the assignments in this document were written by me. The authors of each assignment (Professors at USC) will be listed before each section.

MATH 425A: Fundamental Concepts of Analysis

All assignments in this section were written by Masoud Zargar, RTPC Assistant Professor of Mathematics, USC. Solutions to assignments 1 through 3 are provided.

Assignment 1

Problem 1

(5 points). Differentiate the following functions:

(i) $f(x) = \ln(\sin^2(x))$

(ii) $k(x) = \ln|\cos(\ln x)|$

Solution

(i) Applying the chain rule once yields

$$\frac{d}{dx}f(x) = \frac{d}{dx}ln(sin^2(x)) = \frac{1}{sin^2(x)}\frac{d}{dx}sin^2(x) \qquad (1)$$

Applying the chain rule to $sin^2(x)$ yields

$$\frac{d}{dx}sin^2(x) = 2sin(x)\frac{d}{dx}sin(x) = 2sin(x)cos(x) \qquad (2)$$

Plugging the result from (2) into (1) yields

$$\frac{d}{dx}f(x) = \frac{1}{sin^2(x)} \cdot 2 \cdot sin(x) \cdot cos(x) = 2\frac{cos(x)}{sin(x)} = 2cot(x) \qquad (3)$$

From (3), we get the conclusion that $\frac{d}{dx}f(x) = 2cot(x)$.

(ii) Applying the chain rule once yields

$$\frac{d}{dx}k(x) = \frac{d}{dx}ln|cos(ln(x))| = \frac{1}{cos(ln(x))}\frac{d}{dx}cos(ln(x)) \qquad (4)$$

Applying the chain rule to $cos(ln(x))$ yields

$$\frac{d}{dx}cos(ln(x)) = -sin(ln(x)) \cdot \frac{d}{dx}ln(x) = -\frac{sin(ln(x))}{x} \qquad (5)$$

Plugging the result from (5) into (4) yields

$$\frac{d}{dx}k(x) = -\frac{sin(ln(x))}{xcos(ln(x))} = -\frac{tan(ln(x))}{x} \qquad (6)$$

From (6), we arrive at the conclusion that $\frac{d}{dx}k(x) = -\frac{tan(ln(x))}{x}$.

Problem 2

(5 points). Find f if $f"(x) = x^{-2}$, $x > 0$, $f(1) = f(2) = 0$.

Solution
Applying the Fundamental Theorem of Calculus, integrating $f"(x)$ once with respect to x yields

$$f'(x) = \int f"(x)dx = \int x^{-2}dx = -\frac{1}{x} + C \qquad (7)$$

Integrating the result from (7) with respect to x yields

$$f(x) = \int f'(x)dx = \int (-\frac{1}{x} + C)dx = -ln|x| + Cx + D \qquad (8)$$

The result from (8) allows us to establish the following system of two equations:

1. $-ln|1| + C + D = C + D = 0$
2. $-ln|2| + 2C + D = 0$

Solving directly, we find

$$C + D = 2C + D - ln(2) \implies C = ln(2) \qquad (9)$$

Plugging the result from (9) back into either equation from the system yields

$$ln(2) + D = 0 \implies D = -ln(2) \qquad (10)$$

Combining the results from (9) and (10) with the result from (8), we find

$$f(x) = -ln|x| + ln(2)x - ln(2) \qquad (11)$$

which completes the problem.

Problem 3

(5 points). Compute the limit

$$\lim_{x\to 1}\frac{2^{cos^2(\pi x)+2}-2^{3x}}{sin(\pi x)}$$

Solution If we try to evaluate the limit directly, we find

$$\left.\frac{2^{cos^2(\pi x)+2}-2^{3x}}{sin(\pi x)}\right|_{x=1}=\frac{2^{cos^2(\pi)+1}-2^3}{sin(\pi)}=\frac{2^3-2^3}{0}=\frac{0}{0} \qquad (12)$$

The result from (12) is undefined, but its form tells us we can apply L'Hopital's rule to find

$$\lim_{x\to 1}\frac{2^{cos^2(\pi x)+2}-2^{3x}}{sin(\pi x)}=\lim_{x\to 1}\frac{\frac{d}{dx}(2^{cos^2(\pi x)+2}-2^{3x})}{\frac{d}{dx}sin(\pi x)} \qquad (13)$$

We can quickly compute that the derivative in the denominator evaluates to

$$\frac{d}{dx}sin(\pi x)=cos(\pi x)\frac{d}{dx}(\pi x)=\pi cos(\pi x) \qquad (14)$$

For the derivative in the numerator, note that

$$\frac{d}{dx}(2^{cos^2(\pi x)+2}-2^{3x})=\frac{d}{dx}(2^{cos^2(\pi x)+2})-\frac{d}{dx}2^{3x} \qquad (15)$$

Let $h_1(x):=2^{cos^2(\pi x)+2}$ and $h_2(x):=2^{3x}$ so that $\frac{d}{dx}(2^{cos^2(\pi x)+2}-2^{3x})=\frac{d}{dx}h_1(x)-\frac{d}{dx}h_2(x)$. Then

$$ln(h_1(x))=(cos^2(\pi x)+2)ln(2) \qquad (16)$$

and

$$ln(h_2(x))=3xln(2) \qquad (17)$$

Differentiating both sides of (16) and (17) with respect to x yields

$$\frac{\frac{d}{dx}h_1(x)}{h_1(x)}=\frac{d}{dx}(cos^2(\pi x)+2)ln(2)=ln(2)\frac{d}{dx}cos^2(\pi x)=-2\pi ln(2)cos(\pi x)sin(\pi x) \qquad (18)$$

and

$$\frac{\frac{d}{dx}h_2(x)}{h_2(x)}=\frac{d}{dx}3xln(2)=3ln(2) \qquad (19)$$

Rearranging the results from (18) and (19), we find

$$\frac{d}{dx}h_1(x)=-2\pi ln(2)cos(\pi x)sin(\pi x)2^{cos^2(\pi x)+2} \qquad (20)$$

and

$$\frac{d}{dx}h_2(x)=3ln(2)2^{3x} \qquad (21)$$

Combining (20) and (21), we find

$$\frac{d}{dx}(2^{cos^2(\pi x)+2}-2^{3x})=-2\pi ln(2)cos(\pi x)sin(\pi x)2^{cos^2(\pi x)+2}-3ln(2)2^{3x} \qquad (22)$$

Plugging (22) and (14) into (13) yields

$$\lim_{x\to 1}\frac{2^{cos^2(\pi x)+2}-2^{3x}}{sin(\pi x)}=\lim_{x\to 1}\frac{-2\pi ln(2)cos(\pi x)sin(\pi x)2^{cos^2(\pi x)+2}-3ln(2)2^{3x}}{\pi cos(\pi x)}=\frac{0-24ln(2)}{-\pi}=\frac{24ln(2)}{\pi} \qquad (23)$$

From (23), we conclude that

$$\lim_{x\to 1}\frac{2^{cos^2(\pi x)+2}-2^{3x}}{sin(\pi x)}=\frac{24ln(2)}{\pi} \qquad (24)$$

which completes the problem.

Problem 4

(5 points). Compute the limit

$$\lim_{x\to 0^+}(tan(ex))^x$$

Solution

Since e^x and $ln(x)$ are inverses of each other, both of which are continuous over all of R^+, we have

$$\lim_{x\to 0^+}(tan(ex))^x = \lim_{x\to 0^+} e^{ln((tan(ex))^x)} = \lim_{x\to 0^+} e^{xln(tan(ex))} = exp(\lim_{x\to 0^+} xln(tan(ex))) \tag{25}$$

Directly evaluating the limit from the RHS of (25) yields

$$xln(tan(ex))|_{x=0} = 0\cdot -\infty = -\frac{\infty}{\infty} \tag{26}$$

The result from (26) is undefined, but its form allows us to apply L'Hopital's rule to find

$$\lim_{x\to 0^+} xln(tan(ex)) = \lim_{x\to 0^+} \frac{ln(tan(ex))}{\frac{1}{x}} = \lim_{x\to 0^+} \frac{\frac{d}{dx}ln(tan(ex))}{\frac{d}{dx}\frac{1}{x}} \tag{27}$$

Differentiating the numerator and denominator from the RHS of (27) with respect to x yields

$$\frac{d}{dx}ln(tan(ex)) = \frac{1}{tan(ex)}\cdot sec^2(ex)\cdot e = e\frac{cos(ex)}{sin(ex)}\cdot\frac{1}{cos^2(ex)} = \frac{e}{sin(ex)cos(ex)} \tag{28}$$

and

$$\frac{d}{dx}\frac{1}{x} = \frac{d}{dx}x^{-1} = -x^{-2} \tag{29}$$

Plugging the results from (28) and (29) into (27) yields

$$\lim_{x\to 0^+} xln(tan(ex)) = \lim_{x\to 0^+} \frac{\frac{e}{sin(ex)cos(ex)}}{-x^{-2}} = \lim_{x\to 0^+} \frac{-ex^2}{sin(ex)cos(ex)} = -2e\lim_{x\to 0^+}\frac{x^2}{sin(2ex)} \tag{30}$$

with the last equality following since $2sin(x)cos(x) = sin(2x) \implies \frac{1}{2}sin(2ex) = sin(ex)cos(ex)$. If we try to evaluate the limit from (30) directly, we will get $\frac{0}{0}$ since $\lim_{x\to 0^+} x^2 = 0$ and $\lim_{x\to 0^+} sin(2ex) = 0$. Thus, we apply L'Hopital's rule once more to find

$$\lim_{x\to 0^+} xln(tan(ex)) = \lim_{x\to 0^+} \frac{\frac{d}{dx}x^2}{\frac{d}{dx}sin(2ex)} = \lim_{x\to 0^+}\frac{2x}{2ecos(2ex)} = \frac{0}{2ecos(0)} = \frac{0}{2e} = 0 \tag{31}$$

Plugging the result from (31) into (25), we find

$$\lim_{x\to 0^+}(tan(ex))^x = exp\Big(\lim_{x\to 0^+} xln(tan(ex))\Big) = e^0 = 1 \tag{32}$$

From (32), we conclude that

$$\lim_{x\to 0^+}(tan(ex))^x = 1 \tag{33}$$

which completes the problem.

Problem 5

(5 points). Compute $\int x^2 arctan(x)dx$.

Solution

We will integrate by parts. First, let $u = x^2$ and $dv = arctan(x)dx$ so that $du = 2xdx$ and $v = \int arctan(x)dx$. To compute v explicitly, we must do another integration by parts. Let $u_1 = arctan(x)$ and $dv_1 = dx$ so that $du_1 = \frac{1}{1+x^2}$ and $v_1 = x$. Then

$$\int arctan(x)dx = xarctan(x) - \int \frac{x}{1+x^2}dx \qquad (34)$$

Let $u_2 = 1 + x^2$ so $du_2 = 2xdx$. Then

$$\int \frac{x}{1+x^2}dx = \frac{1}{2}\int \frac{1}{u_2}du_2 = \frac{1}{2}ln|u_2| + C = \frac{1}{2}ln|1+x^2| + C \qquad (35)$$

Choosing $C = 0$ and plugging the result from (35) into (34), we find

$$v := \int arctan(x)dx = xarctan(x) - \frac{1}{2}ln|1+x^2| \qquad (36)$$

Now that we have u, v, du, and dv explicitly defined, we can integrate by parts to find

$$\int x^2 arctan(x)dx = x^3 arctan(x) - \frac{x^2}{2}ln|1+x^2| - 2\int x^2 arctan(x)dx + \int xln|1+x^2|dx \qquad (37)$$

Combining like terms from (37) and using the definition of u_2 as before, we find

$$3\int x^2 arctan(x)dx = x^3 arctan(x) - \frac{x^2}{2}ln|1+x^2| + \frac{1}{2}\int ln|u_2|du_2 \qquad (38)$$

To evaluate the integral from the RHS of (38), we need to integrate by parts again. Let $u_3 = ln|u_2|$ and $dv_3 = du_2$ so $du_3 = \frac{du_2}{u_2}$ and $v_3 = u_2$. Integrating by parts yields

$$\int ln|u_2|du_2 = u_2 ln|u_2| - \int \frac{u_2}{u_2}du_2 = u_2 ln|u_2| - \int du_2 = u_2 ln|u_2| - u_2 \qquad (39)$$

Plugging the result from (38) into (39) and writing the result in terms of x yields

$$\begin{aligned}
3\int x^2 arctan(x)dx = &\quad x^3 arctan(x) - \frac{x^2}{2}ln|1+x^2| + \frac{1}{2}(u_2 ln|u_2| - u_2) \\
= &\quad x^3 arctan(x) - \frac{x^2}{2}ln|1+x^2| + \frac{1}{2}((1+x^2)ln(1+x^2) - x^2 - 1) \\
= &\quad x^3 arctan(x) - \frac{x^2}{2}ln|1+x^2| + \frac{1}{2}ln(1+x^2) + \frac{x^2}{2}ln(1+x^2) - \frac{x^2}{2} - \frac{1}{2} \\
= &\quad x^3 arctan(x) + \frac{1}{2}ln|1+x^2| - \frac{x^2}{2} - \frac{1}{2} \qquad (40)
\end{aligned}$$

Dividing both sides of (40) by three yields

$$\int x^2 arctan(x)dx = \frac{2x^3 arctan(x) + ln(1+x^2) - x^2 - 1}{6} + C = \frac{2x^3 arctan(x) + ln(1+x^2) - x^2}{6} + C \qquad (41)$$

which completes the problem.

Problem 6

(5 points). Compute the indefinite integral

$$\int \frac{(lnx)^3 dx}{x\sqrt{1-(lnx)^2}}$$

Solution

We will use trigonometric substitution. Let $ln(x) = sin(u)$ so $\frac{dx}{x} = cos(u)du$ and $u = arcsin(ln(x))$. Then

$$\int \frac{(lnx)^3 dx}{x\sqrt{1-(lnx)^2}} = \int \frac{sin^3(u)cos(u)du}{\sqrt{1-sin^2(u)}} = \int \frac{sin^3(u)cos(u)du}{\sqrt{cos^2(u)}} = \int sin^3(u)du \tag{42}$$

Utilizing the fact that $sin^2(x) + cos^2(x) = 1$ for all $x \in \mathbb{R}$, we find

$$\int \frac{(lnx)^3 dx}{x\sqrt{1-(lnx)^2}} = \int (1-cos^2(u))sin(u)du = \int sin(u)du - \int cos^2(u)sin(u)du = -cos(u) - \int cos^2(u)sin(u)du \tag{43}$$

Now, let $v = cos(u)$ so that $dv = -sin(u)du$, and we find

$$\int \frac{(lnx)^3 dx}{x\sqrt{1-(lnx)^2}} = -cos(u) + \int v^2 dv = -cos(u) + \frac{v^3}{3} = \frac{cos^3(u)}{3} - cos(u) \tag{44}$$

Expressing (44) in terms of x yields

$$\int \frac{(lnx)^3 dx}{x\sqrt{1-(lnx)^2}} = \frac{cos^3(arcsin(ln(x)))}{3} - cos(arcsin(ln(x))) \tag{45}$$

Note that

$$cos(arcsin(ln(x)) = \sqrt{1-sin^2(arcsin(ln(x))} = \sqrt{1-ln^2(x)} \tag{46}$$

Plugging the result from (46) into (45) yields

$$\int \frac{(lnx)^3 dx}{x\sqrt{1-(lnx)^2}} = \frac{(1-ln^2(x))^{\frac{3}{2}}}{3} - \sqrt{1-ln^2(x)} \tag{47}$$

The result from (47) lets us conclude that

$$\int \frac{(lnx)^3 dx}{x\sqrt{1-(lnx)^2}} = \frac{(1-ln^2(x))^{\frac{3}{2}}}{3} - \sqrt{1-ln^2(x)} + C \tag{48}$$

which completes the problem.

Problem 7

(5 points). Compute

$$\lim_{x\to\infty} e^{-x^2}\int_x^{x+\frac{ln(x)}{x}} e^{t^2}dt$$

Solution

First, rewrite

$$\lim_{x\to\infty} e^{-x^2}\int_x^{x+\frac{ln(x)}{x}} e^{t^2}dt = \lim_{x\to\infty}\frac{\int_x^{x+\frac{ln(x)}{x}} e^{t^2}dt}{e^{x^2}} \qquad (49)$$

and note that

$$\lim_{x\to\infty} e^{x^2} = \infty \qquad (50)$$

Next, note that, for all $t \in [x, x+\frac{ln(x)}{x}]$, $e^{t^2} \geq e^{x^2}$, so we know

$$\lim_{x\to\infty}\int_x^{x+\frac{ln(x)}{x}} e^{t^2}dt \geq \lim_{x\to\infty} e^{x^2}\int_x^{x+\frac{ln(x)}{x}} dt = \lim_{x\to\infty} e^{x^2}(x+\frac{ln(x)}{x}-x) = \lim_{x\to\infty} e^{x^2}\frac{ln(x)}{x} \qquad (51)$$

Claim: $e^x \geq x$ for all $x \in \mathbb{R}$ such that $x \geq 0$.
Proof. For $x = 0$, we have $e^x = e^0 = 1 \geq 0 = x$. Thus, it suffices to show e^x increases at least as fast as x for all $x \geq 0$. Note that $\frac{d}{dx}e^x = e^x$ and $\frac{d}{dx}x = 1$, so it suffices to show $e^x \geq 1$ for all $x \geq 0$. For $x = 0$, $e^x = e^0 = 1 \geq 1$. Also, $\frac{d^2}{dx^2}e^x = e^x > 0$ for all $x \in \mathbb{R}$ by definition of the exponential function. Thus, the rate of change of $\frac{d}{dx}e^x = e^x$ is *strictly* positive for all $x \in \mathbb{R}$, so $e^x > e^0$ for all $x > 0$. This combines with the fact that $e^0 \geq 1$ to prove that $e^x \geq 1$ for all $x \geq 0$. Thus, $e^x \geq x$ when $x = 0$, and e^x grows at least as fast as x for all $x \geq 0$, so $e^x \geq x$ for all $x \geq 0$.
From this result, substituting x^2 for x, we find

$$lim_{x\to\infty}\int_x^{x+\frac{ln(x)}{x}} e^{t^2}dt \geq \lim_{x\to\infty} x^2\frac{ln(x)}{x} = \lim_{x\to\infty} xln(x) = \infty \qquad (52)$$

From (50) and (52), we see that directly evaluating

$$\lim_{x\to\infty}\frac{\int_x^{x+\frac{ln(x)}{x}} e^{t^2}dt}{e^{x^2}}$$

leads to $\frac{\infty}{\infty}$, so we apply L'Hopital's rule to find

$$\lim_{x\to\infty}\frac{\int_x^{x+\frac{ln(x)}{x}} e^{t^2}dt}{e^{x^2}} = \lim_{x\to\infty}\frac{\frac{d}{dx}\int_x^{x+\frac{ln(x)}{x}} e^{t^2}dt}{\frac{d}{dx}e^{x^2}} \qquad (53)$$

We can directly evaluate the derivative in the denominator to find

$$\frac{d}{dx}e^{x^2} = 2xe^{x^2} \qquad (54)$$

For the derivative in the numerator, we use the Fundamental Theorem of Calculus in conjunction with the chain rule to find

$$\frac{d}{dx}\int_x^{x+\frac{ln(x)}{x}} e^{t^2}dt = e^{(x+\frac{ln(x)}{x})^2}\frac{d}{dx}(x+\frac{ln(x)}{x}) - e^{x^2} \qquad (55)$$

Applying the quotient rule to the derivative from (55), we find

$$\frac{d}{dx}(x+\frac{ln(x)}{x}) = 1+\frac{1-ln(x)}{x^2} \qquad (56)$$

Plugging the result from (56) into (55) and simplifying yields

$$\begin{aligned}\frac{d}{dx}\int_x^{x+\frac{ln(x)}{x}} e^{t^2}dt = &\quad e^{(x+\frac{ln(x)}{x})^2}(1+\frac{1-ln(x)}{x^2}) - e^{x^2} = e^{x^2}(e^{2ln(x)}e^{(\frac{ln(x)}{x})^2}(1+\frac{1-ln(x)}{x^2}) - 1)\\ = &\quad e^{x^2}(x^2e^{(\frac{ln(x)}{x})^2}(1+\frac{1-ln(x)}{x^2}) - 1) \qquad (57)\end{aligned}$$

Plugging the results from (57) and (54) into (53), we find

$$\lim_{x\to\infty}\frac{\int_x^{x+\frac{ln(x)}{x}} e^{t^2}dt}{e^{x^2}} = \lim_{x\to\infty}\frac{e^{x^2}(x^2e^{(\frac{ln(x)}{x})^2}(1+\frac{1-ln(x)}{x^2}) - 1)}{2xe^{x^2}} = \lim_{x\to\infty}\frac{(x^2e^{(\frac{ln(x)}{x})^2}(1+\frac{1-ln(x)}{x^2}) - 1)}{2x} \qquad (58)$$

Directly evaluating the limit from the RHS of (58) yields

$$\lim_{x\to\infty}\frac{(x^2e^{(\frac{ln(x)}{x})^2}(1+\frac{1-ln(x)}{x^2}) - 1)}{2x} = \lim_{x\to\infty}\frac{x^2-1}{2x} \qquad (59)$$

since $\lim_{x\to\infty}\frac{ln(x)}{x} = 0$ and $\lim_{x\to\infty}\frac{1-ln(x)}{x^2} = 0$. Evaluating (59) directly leads to $\frac{\infty}{\infty}$, so we apply L'Hopital's rule once more to find

$$\lim_{x\to\infty}\frac{\int_x^{x+\frac{ln(x)}{x}} e^{t^2}dt}{e^{x^2}} = \lim_{x\to\infty}\frac{\frac{d}{dx}x^2-1}{\frac{d}{dx}2x} = \lim_{x\to\infty}\frac{2x}{2} = \lim_{x\to\infty}x = \infty \qquad (60)$$

From (60), we conclude that

$$\lim_{x\to\infty} e^{-x^2}\int_x^{x+\frac{ln(x)}{x}} e^{t^2}dt = \infty$$

which completes the problem.

Assignment 2

Problem 1

(5 points). Suppose $f : X \to Y$ is a function, and that $\{A_i : i \in I\} \subseteq \mathcal{P}(X)$ and $\{B_j : j \in J\} \subseteq \mathcal{P}(Y)$. Rigorously prove the following:

(i) $f(\bigcup_{i\in I} A_i) = \bigcup_{i\in I} f(A_i)$,

(ii) $f^{-1}(\bigcup_{j\in J} B_j) = \bigcup_{j\in J} f^{-1}(B_j)$,

(iii) $f^{-1}(Y \setminus B) = X \setminus f^{-1}(B)$ for every subset $B \subseteq Y$.

Solution

(i) By the axiom of extensionality, it suffices to show

$$\forall y((y \in f(\bigcup_{i\in I} A_i)) \iff (y \in \bigcup_{i\in I} f(A_i))$$

Consider an arbitrary y. By the definition of the image,

$$y \in f(\bigcup_{i\in I} A_i) \implies \exists x \in \bigcup_{i\in I} A_i \text{ s.t. } y = f(x)$$

By an inductive application of the definition of the union of two sets,

$$\exists x \in \bigcup_{i\in I} A_i \text{ s.t. } y = f(x) \implies \exists x \in X \text{ s.t. } (y = f(x)) \wedge (x \in A_i) \text{ for some } i \in I$$

Since $\{A_i : i \in I\} \subseteq \mathcal{P}(X)$, $A_i \subseteq X$ for all $i \in I$, so $x \in A_i \implies x \in X$ for all $i \in I$. Thus, we can write

$$\exists x \in X \text{ s.t. } (y = f(x)) \wedge (x \in A_i) \text{ for some } i \in I \implies \exists x \in A_i \text{ s.t. } y = f(x) \text{ for some } i \in I$$

The definition of the image yields

$$\exists x \in A_i, i \in I \text{ s.t. } y = f(x) \implies \exists i \in I \text{ s.t. } y \in f(A_i)$$

Finally, and inductive application of the definition of the union of two sets yields

$$\exists i \in I \text{ s.t. } y \in f(A_i) \implies y \in \bigcup_{i\in I} f(A_i)$$

which allows us to conclude

$$y \in f(\bigcup_{i\in I} A_i) \implies y \in \bigcup_{i\in I} f(A_i)$$

Going in the opposite direction, we find

$$\begin{aligned} y \in \bigcup_{i\in I} f(A_i) &\implies && \exists i \in I \text{ s.t. } y \in f(A_i) \\ &\implies && \exists x \in A_i \text{ s.t. } y = f(x) \text{ for some } i \in I \\ &\implies && \exists x \in \bigcup_{i\in I} A_i \text{ s.t. } y = f(x) \\ &\implies && y \in f(\bigcup_{i\in I} A_i) \end{aligned}$$

where the first and third implications follow by an inductive application of the definition of the union of two sets and the second and fourth implications follow by the definition of the image. Thus,

$$\forall y((y \in f(\bigcup_{i\in I} A_i)) \iff (y \in \bigcup_{i\in I} f(A_i)))$$

which completes the proof that

$$f(\bigcup_{i\in I} A_i) = \bigcup_{i\in I} f(A_i)$$

(ii) By the axiom of extensionality, it suffices to show

$$\forall x((x \in f^{-1}(\bigcup_{j\in J} B_j)) \iff (x \in \bigcup_{j\in J} f^{-1}(B_j)))$$

Consider an arbitrary x. By the definition of the pre-image,

$$x \in f^{-1}(\bigcup_{j\in J} B_j) \implies f(x) \in \bigcup_{j\in J} B_j$$

Inductively applying the definition of the union of two sets, we find

$$f(x) \in \bigcup_{j\in J} B_j \implies \exists j \in J \text{ s.t. } f(x) \in B_j$$

By the definition of the pre-image, we have

$$\exists j \in J \text{ s.t. } f(x) \in B_j \implies \exists j \in J \text{ s.t. } x \in f^{-1}(B_j)$$

Via one more inductive application of the definition of the union of two sets, we find

$$\exists j \in J \text{ s.t. } x \in f^{-1}(B_j) \implies x \in \bigcup_{j\in J} f^{-1}(B_j)$$

which allows us to conclude

$$x \in f^{-1}(\bigcup_{j\in J} B_j) \implies x \in \bigcup_{j\in J} f^{-1}(B_j)$$

Going in the opposite direction, we find

$$\begin{aligned}
x \in \bigcup_{j\in J} f^{-1}(B_j) &\implies && \exists j \in J \text{ s.t. } x \in f^{-1}(B_j) \\
&\implies && \exists j \in J \text{ s.t. } f(x) \in B_j \\
&\implies && f(x) \in \bigcup_{j\in J} B_j \\
&\implies && x \in f^{-1}(\bigcup_{j\in J} B_j)
\end{aligned}$$

where the first and third implications follow by inductive applications of the definition of the union of two sets, and the second and fourth follow by the definition of the pre-image. Thus,

$$\forall x((x \in f^{-1}(\bigcup_{j\in J} B_j)) \iff (x \in \bigcup_{j\in J} f^{-1}(B_j)))$$

which completes the proof that

$$f^{-1}(\bigcup_{j\in J} B_j) = \bigcup_{j\in J} f^{-1}(B_j)$$

(iii) By the axiom of extensionality, it suffices to show

$$\forall x((x \in f^{-1}(Y \setminus B)) \iff (x \in X \setminus f^{-1}(B)))$$

for an arbitrary subset $B \subseteq Y$. Consider an arbitrary x and an arbitrary $B \subseteq Y$. By definition of the pre-image, we have

$$x \in f^{-1}(Y \setminus B) \implies f(x) \in Y \setminus B$$

By definition of the difference of two sets,

$$f(x) \in Y \setminus B \implies (f(x) \in Y) \wedge (f(x) \notin B)$$

Applying the definition of the pre-image twice yields

$$(f(x) \in Y) \wedge (f(x) \notin B) \implies (x \in f^{-1}(Y)) \wedge (f \notin f^{-1}(B))$$

Since $f : X \to Y$, we know $f^{-1}(Y) = X$, so we can simplify the above implication to

$$(f(x) \in Y) \wedge (f(x) \notin B) \implies (x \in X) \wedge (x \notin f^{-1}(B))$$

Applying the definition of the difference of two sets once more yields

$$x \in X \setminus f^{-1}(B)$$

which allows us to conclude

$$x \in f^{-1}(Y \setminus B) \implies x \in X \setminus f^{-1}(B)$$

Going in the other direction, we find

$$\begin{aligned}
x \in X \setminus f^{-1}(B) &\implies && (x \in X) \wedge (x \notin f^{-1}(B)) \\
&\equiv && (x \in f^{-1}(Y)) \wedge (x \notin f^{-1}(B)) \\
&\implies && (f(x) \in Y) \wedge (f(x) \notin B) \\
&\implies && f(x) \in Y \setminus B \\
&\implies && x \in f^{-1}(Y \setminus B)
\end{aligned}$$

where the equivalence follows from the previously noted identity $X = f^{-1}(Y)$, the first and third implications follow from the definition of the difference of two sets, and the second and fourth implications follow from the definition of the pre-image. Thus,

$$\forall x((x \in f^{-1}(Y \setminus B)) \iff (x \in X \setminus f^{-1}(B)))$$

for all $B \subseteq Y$, which completes the proof that

$$f^{-1}(Y \setminus B) = X \setminus f^{-1}(B) \text{ for every subset } B \subseteq Y$$

and thus completes the problem.

Problem 2

(5 points). Recall that $Y^X := Fun(X, Y) = \{f : X \to Y\}$ is the set of functions from X to Y. Suppose X, Y, Z are three sets. Prove that $|(Z^Y)^X| = |Z^{X \times Y}|$ (note: when the sets are finite, this reduces to a well-known algebraic identity, well-known to infants) by constructing

$$\Phi : Fun(X, Fun(Y, Z)) \to Fun(X \times Y, Z)$$

and

$$\Psi : Fun(X \times Y, Z) \to Fun(X, Fun(Y, Z))$$

such that both $\Phi \circ \Psi$ and $\Psi \circ \Phi$ are identity. functions. Why does this prove the equality of cardinalities?

Solution
For any $f^* \in (Z^Y)^X$, define

$$\Phi(f^*) = f \in Z^{X \times Y} \text{ s.t. } f(x,y) = (f^*(x))(y)$$

for all $x \in X$, $y \in Y$, $(x,y) \in X \times Y$. Here $(f^*(x))(y)$ denotes the function $f^*(x) : Y \to Z$ evaluated at $y \in Y$. That is,

$$(\Phi(f^*))(x,y) = (f^*(x))(y)$$

for all $x \in X$, $y \in Y$, $(x,y) \in X \times Y$, and $f^* \in (Z^Y)^X$.
Similarly, for any $f^+ \in Z^{X \times Y}$ define

$$\Psi(f^+) = f \in (Z^Y)^X \text{ s.t. } (f(x))(y) = f^+(x,y)$$

for all $x \in X$, $y \in Y$, $(x,y) \in X \times Y$. Here $(f(x))(y)$ denotes the function $f(x) : Y \to Z$ evaluated at $y \in Y$. That is,

$$((\Psi(f^+))(x))(y) = f^+(x,y)$$

for all $x \in X$, $y \in Y$, $(x,y) \in X \times Y$, and $f^+ \in Z^{X \times Y}$
For any $f^* \in (Z^Y)^X$, note that

$$\begin{aligned}
\Psi \circ \Phi(f^*) = &\quad \Psi(f' \in Z^{X \times Y} \text{ s.t. } f'(x,y) = (f^*(x))(y) \quad \forall x \in X, y \in Y, (x,y) \in X \times Y) \\
= &\quad f \in (Z^Y)^X \text{ s.t. } (f(x))(y) = f'(x,y) \quad \forall x \in X, y \in Y, (x,y) \in X \times Y \\
= &\quad f \in (Z^Y)^X \text{ s.t. } (f(x))(y) = (f^*(x))(y) \quad \forall x \in X, y \in Y, (x,y) \in X \times Y \\
= &\quad f^*
\end{aligned}$$

so $\Psi \circ \Phi = id_{(Z^Y)^X}$.
Similarly, for any $f^+ \in Z^{X \times Y}$, note that

$$\begin{aligned}
\Phi \circ \Psi(f^+) = &\quad \Phi(f' \in (Z^Y)^X \text{ s.t. } (f'(x))(y) = f^+(x,y) \quad \forall x \in X, y \in Y, (x,y) \in X \times Y) \\
= &\quad f \in Z^{X \times Y} \text{ s.t. } f(x,y) = (f'(x))(y) \quad \forall x \in X, y \in Y, (x,y) \in X \times Y) \\
= &\quad f \in Z^{X \times Y} \text{ s.t. } f(x,y) = f^+(x,y) \forall x \in X, y \in Y, (x,y) \in X \times Y) \\
= &\quad f^+
\end{aligned}$$

so $\Phi \circ \Psi = id_{Z^{X \times Y}}$. Since

$$\Psi \circ \Phi = id_{(Z^Y)^X} \qquad \text{and} \qquad \Phi \circ \Psi = id_{Z^{X \times Y}}$$

we know from lecture that Φ is a bijection from $(Z^Y)^X$ to $Z^{X \times Y}$. By the definition of equality of cardinalities, this implies

$$|(Z^Y)^X| = |Z^{X \times Y}|$$

which completes the proof.

Problem 3

(5 points). Suppose $\Phi : C^0(\mathbb{R};\mathbb{R}) \to \mathbb{R}_{\geq 0}$ is a non-negative function on the set $C^0(R;R)$ of continuous functions $\mathbb{R} \to \mathbb{R}$. Suppose further that there is a constant $B > 0$ such that for every $k \in \mathbb{N}$ and every k *distinct* continuous functions $f_1, ..., f_k \in C^0(\mathbb{R};\mathbb{R})$, we have

$$\sum_{i=1}^{k} \Phi(f_i) \leq B.$$

Prove that $\{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \neq 0\}$ is countable. Can you generalize this problem?

Hint: If $a > 0$ is a real number, note that there is an $n \in N$ such that $a \geq \frac{1}{n}$. Look at the proof of the countability of algebraic numbers for inspiration.

Solution

Since $\Phi(f) \geq 0, \in \mathbb{R}$ for all $f \in C^0(\mathbb{R};\mathbb{R})$ by definition, we know $\Phi(f) \neq 0 \implies \Phi(f) > 0$ for all $f \in C^0(\mathbb{R};\mathbb{R})$. That is,

$$f \in \{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \neq 0\} \implies \Phi(f) \in R_+$$

Following the hint, note that for any such $\Phi(f) \in \mathbb{R}_+$, we can find an $n \in \mathbb{N}$ such that $\Phi(f) \geq \frac{1}{n}$. Thus,

$$\begin{aligned} f \in \{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \neq 0\} &\implies \quad f \in \{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \geq \frac{1}{n} \text{ for some } n \in \mathbb{N}\} \\ &\implies \quad f \in \bigcup_{n\in\mathbb{N}} \{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \geq \frac{1}{n}\} \end{aligned}$$

for all $f \in \{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \neq 0\}$ and

$$\begin{aligned} f \in \bigcup_{n\in\mathbb{N}} \{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \geq \frac{1}{n}\} &\implies \quad f \in \{f \in C^0(R;R) : \Phi(f) > 0\} \\ &\implies \quad f \in \{f \in C^0(R;R) : \Phi(f) \neq 0\} \end{aligned}$$

for all $f \in \{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \neq 0\}$. Thus, we have

$$\forall f((f \in \{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \neq 0\}) \iff (f \in \bigcup_{n\in N} \{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \geq \frac{1}{n}\}))$$

so

$$\{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \neq 0\} = \bigcup_{n\in N} \{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \geq \frac{1}{n}\}$$

We claim that $\{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \geq \frac{1}{n}\}$ is finite for all $n \in \mathbb{N}$.
Assume to the contrary that $\{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \geq \frac{1}{n}\}$ is infinite for some $n \in \mathbb{N}$. Then we can choose $n\lceil B\rceil + 1 > nB$ of its elements

$$f_1, ..., f_{n\lceil B\rceil+1} \in \{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \geq \frac{1}{n}\}$$

and

$$\sum_{i=1}^{n\lceil B\rceil+1} \Phi(f_i) \geq \sum_{i=1}^{n\lceil B\rceil+1} \frac{1}{n} = (n\lceil B\rceil + 1)\frac{1}{n} > n\lceil B\rceil\frac{1}{n} = B$$

This directly contradicts the upper bound on $\sum_{i=1}^{k} \Phi(f_i)$ for distinct $f_1, ..., f_k \in C^0(\mathbb{R};\mathbb{R})$ from the problem statement. Thus,

$$\{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \geq \frac{1}{n}\}$$

is finite (and thus countable) for all $n \in \mathbb{N}$. Since $\mathbb{N}$ is countable by definition, $\{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \neq 0\}$ is thus a countable union of countable sets, so it too is countable by Proposition 1.5 from the notes. This completes the proof that $\{f \in C^0(\mathbb{R};\mathbb{R}) : \Phi(f) \neq 0\}$ is countable.

To generalize this problem, consider an arbitrary $\Phi : X \to Y$ such that $\Phi(x) \geq 0$ for all $x \in X$ and for all $k \in \mathbb{N}$ and distinct $x_1, .., x_k \in X$,

$$\sum_{i=1}^{k} \Phi(x) \leq B$$

for some constant B. Then

$$\{x \in X : \Phi(x) \neq 0\} = \bigcup_{n \in \mathbb{N}} \{x \in X : x \geq \frac{1}{n}\}$$

by the exact same logic as before. Moreover, $\{x \in X : \Phi(x) \neq 0\}$ a countable union of countable sets, so it is also countable by Proposition 1.5. This generalization completes the problem.

Problem 4

(5 points). In the multiverse/universe **Infinitum 425a**, at every point in $\mathbb{R}^3$, there is a (point) planet, and there is a two-way/undirected road between $x, y \in \mathbb{R}^3$ if and only if $|x-y| = 1$, that is, they are 1 unit apart. Using Zorn's lemma, prove that there is a way to connect all the planets using some of the roads such that there is *exactly one* path (finitely many roads) between any two planets.

Solution

Let $G_p = (V, E)$ be the graph of **Infinitum 425a**, where

$$V = \mathbb{R}^3$$

is the set of all vertices (planets) and

$$E = \{(x, y) | x, y \in \mathbb{R}^3, |x - y| = 1\}$$

is the set of all edges (roads) between any two points (planets) in $\mathbb{R}^3$ that are exactly 1 unit apart. Note that the two directed edges $(x, y), (y, x) \in E$ provide the functionality of the two-way roads from the problem description.

First, we claim that G_p is connected. To prove this, it suffices to show that, for any two *distinct* points $x, y \in V$ (i.e. $|x - y| \neq 0$), we can find a sequence of edges $e_1, ..., e_k \in E$ that connect x with y. Suppose $|x - y| = D \in \mathbb{R}^+$, and consider the line segment $\overline{xy}$ of length D connecting x and y. If $D > 1$, we can traverse $\overline{xy}$ in 1 unit steps (since there are roads between all $a, b \in \mathbb{R}^3$ such that $|a - b| = 1$) until the remaining distance is $d := D - \lfloor D \rfloor < 1$ where $d \in \mathbb{R}_{\geq 0}$. That is, we can find $e_1, ..., e_{\lfloor D \rfloor} \in E$ that connect x to a point x' which satisfies $|x' - y| < 1$.

If $d = 0$, $x' = y$, and we have connected x to y using $e_1, ..., e_{\lfloor D \rfloor} \in E$.

On the other hand, if $d \in (0, 1)$, then consider the line segment $\overline{x'y}$ of length d connecting x' with y and define x" to be its midpoint. Then $|x' - x"| = |y - x"| = \frac{d}{2}$. Consider an arbitrary line of length $\sqrt{1 - \frac{d^2}{4}}$ orthogonal to $\overline{x'y}$ connecting x" to some other point $x^* \in \mathbb{R}^3$. Note that, since $\overline{x'x"}$ is a sub-segment of $\overline{x'y}$, $\overline{x'y} \perp \overline{x"x^*}$ implies $\overline{x'x^*}$ forms a right triangle with $\overline{x'x"}$ and $\overline{x"x^*}$. Applying the Pythagorean Theorem, we find that

$$|x' - x^*| = \sqrt{(\sqrt{1 - \frac{d^2}{4}})^2 + (\frac{d}{2})^2} = \sqrt{1 - \frac{d^2}{4} + \frac{d^2}{4}} = \sqrt{1} = 1$$

That is, x' and x^* are exactly 1 unit apart, so there is a road $e_{\lfloor D \rfloor + 1} \in E$ directly connecting them. Also, by symmetry, since $\overline{x"y}$ is a sub-segment of $\overline{x'y}$, $\overline{x'y} \perp \overline{x"x^*}$ implies $\overline{x^*y}$ forms a right triangle with $\overline{x"y}$ and $\overline{x"x^*}$. Applying the Pythagorean Theorem, we find that

$$|y - x^*| = \sqrt{(\sqrt{1 - \frac{d^2}{4}})^2 + (\frac{d}{2})^2} = \sqrt{1 - \frac{d^2}{4} + \frac{d^2}{4}} = \sqrt{1} = 1$$

That is, x^* and y are exactly 1 unit apart, so there is a road $e_{\lfloor D \rfloor + 2} \in E$ directly connecting them too. This yields a sequence $e_1, ..., e_{\lfloor D \rfloor}, e_{\lfloor D \rfloor + 1}, e_{\lfloor D \rfloor + 2} \in E$ that connect x and y.

Thus, for any two *distinct* points $x, y \in R^3 = V$, we can find a sequence of edges in E that connect x and y. Since $D := |x - y| < \infty$ is finite, our sequence of $\lfloor D \rfloor + 2$ edges is too. This completes the proof that $G_P := (V, E)$ is connected.

Next, we claim that, for any connected graph $G = (V, E)$ (V and E not necessarily the same as before), there exists a spanning tree M. We will use Zorn's lemma to prove this. Let

$$\Sigma = \{H = (V, E_H) | E_H \subseteq E \text{ and } H \text{ has no cycles}\}$$

be the set of all subgraphs of G with the same vertex set V and without cycles. For any two subgraphs $H, H' \in \Sigma$, we define $\leq$ to be the relation satisfying $H \leq H' \iff H$ is a subgraph of H' (i.e. $\iff E_H \subseteq E_{H'}$ since $V_H = V$ for all $H \in \Sigma$). Note that, for all $H \in \Sigma$,

$$H \leq H$$

since $E_H \subseteq E_H$. Also, for all $H, H' \in \Sigma$,

$$(H \leq H' \wedge H' \leq H) \implies H = H'$$

since

$$H \leq H' \implies E_H \subseteq E_{H'}$$

and

$$H' \leq H \implies E_{H'} \subseteq E_H$$

so

$$(H \leq H' \wedge H' \leq H) \implies (E_H \subseteq E_{H'} \wedge E_{H'} \subseteq E_H) \implies E_H = E_{H'} \implies H = (V, E_H) = (V, E_{H'}) = H'$$

Finally, for all $H, H', H" \in \Sigma$,

$$(H \leq H' \wedge H' \leq H") \implies (H \leq H")$$

since

$$H \leq H' \implies E_H \subseteq E_{H'}$$

and

$$H' \leq H" \implies E_{H'} \subseteq E_{H"}$$

so

$$(H \leq H' \wedge H' \leq H") \implies (E_H \subseteq E_{H'} \subseteq E_{H"}) \implies (E_H \subseteq E_{H"}) \implies (H \leq H")$$

Thus, $(\Sigma, \leq)$ is a pair of a set Σ and a relation $\leq$ which satisfies reflexivity, anti-symmetry, and transitivity, so $(\Sigma, \leq)$ is a poset (partially ordered set).

To apply Zorn's Lemma, we must now prove that, for all chains C, there exists an upper bound $s_C \in \Sigma$. We will do so constructively. We claim that

$$s_C = \bigcup_{H \in C} H$$

is an upper bound for any arbitrary chain C in $(\Sigma, \leq)$. It suffices to show that $s_C \in \Sigma$ and that $H \leq s_C$ for all $H \in C$.
To show $s_C \in \Sigma$, we must show $s_C = (V, E_{s_C})$ where $E_{s_C} \subseteq E$ and s_C has no cycles. Note that, since $H = (V, E_H)$ for all $H \in \Sigma$,

$$s_C = \bigcup_{H \in C} H = (V, \bigcup_{H \in C} E_H)$$

Also, for all $H \in C$, $E_H \subseteq E$, so for all $e \in \bigcup_{H \in C} E_H$, $e \in E$, so $\bigcup_{H \in C} E_H \subseteq E$. Thus

$$s_C = (V, \bigcup_{H \in C} E_H)$$

is a subgraph of G with vertex set V. To show s_C has no cycles, assume to the contrary that s_C has a cycle. Then there are two paths in s_C between $v_1, v_2 \in V$. If these two paths were both present in the same $H \in C$,

then H has a cycle, which contradicts the definition of Σ (since $C \subseteq \Sigma$, so all $H \in C$ must not have cycles). On the other hand, if the two paths were the result of taking the union of some $H_1, H_2 \in C$, where H_1 and H_2 each had at most 1 path between v_1 and v_2, then there must exist some $e_1 \in E_{H_1}$ and some $e_2 \in E_{H_2}$ such that $e_1 \notin E_{H_2}$ and $e_2 \notin E_{H_1}$. But this implies $E_{H_1} \not\subseteq E_{H_2}$ and $E_{H_2} \not\subseteq E_{H_1}$, which implies $H_1 \not\leq H_2$ and $H_2 \not\leq H_1$. This directly contradicts the definition of a chain C, which states that $H_1 \leq H_2$ or $H_2 \leq H_1$ for all $H_1, H_2 \in C$. Thus, by assuming $\bigcup_{H \in C} H$ has a cycle, we have reached a contradiction with the law of excluded middle. This completes the proof that s_C is a subgraph of G with vertex set V and no cycles, so $s_C \in \Sigma$

Now that we have shown $s_C \in \Sigma$ for an arbitrary chain C, we just need to prove H is a subgraph of s_C for all $H \in C$. We already showed that s_C and H have the same vertex set, so it suffices to show $E_H \subseteq E_{s_C}$ for all $H \in C$. For all $H \in C$ and all $e \in E_H$, $e \in \bigcup_{H \in C} E_H = E_{s_C}$, so $E_H \subseteq E_{s_C}$. Thus, for all $H \in C$, we have $V_H = V \subseteq V = V_{s_C}$ and $E_H \subseteq E_{s_C}$. Thus, for all $H \in C$, H is a subgraph of s_C, so $H \leq s_C$. Thus, $s_C := \bigcup H \in C H \in \Sigma$ satisfies

$$H \leq s_C$$

for all $H \in C$, which completes the proof that s_C is an upper bound for any chain C.

By Zorn's Lemma, since every chain C has an upper bound s_C, there exists a maximal element $M \in \Sigma$. That is, for all $H \in \Sigma$,

$$M \leq H \implies M = H$$

Note that, since $M \in \Sigma$, M has V as its vertex set, and M has no cycles.

We claim that M is also connected. Assume to the contrary that M is not connected. Then there exists two points $v_a, v_a \in V$ such that there is no path between v_1 and v_2 in M. However, since G is connected, we know exists a finite path $e_1, .., e_k \in E$ that connects v_a to v_b. Follow this path from v_a to v_b until reaching an edge e_i, connecting vertices v_i and v_{i+1}, such that v_{i+1} is the first vertex along the path which *can not* be reached by v_a only using edges in E_M. That is, v_i is reachable from v_a with edges in E_M, but v_{i+1} is not. Then construct $M' = (V, E_{M'}) = (V, E_M \cup e_i)$. Note that $e_i \notin E_M$, as $e_i \in E_M$ implies v_{i+1} is reachable from v_i which is reachable from v_a using only edges in E_M. Also, $e_i \in E$ and $E_M \subseteq E$, so $E_{M'} \subseteq E$.

We claim that M' has no cycles. Assume to the contrary that M' has a cycle. Since $M \in \Sigma$, M has no cycles, so adding e_i to E_M to construct M' must have created a cycle. This implies that there is exactly one path between v_i and v_{i+1} in M and exactly two paths between v_i and v_{i+1} in M'. However, by definition of e_i, there is a path between v_a and v_i in M, and there is no path betweenv_a and v_{i+1} in M. If there is one path in M between v_i and v_{i+1}, then there is also a path between v_a and v_{i+1} (constructed by appending the path between v_i and v_{i+1} to the path between v_a and v_i), which contradicts the definition of e_i. Thus, adding e_i to E_M to construct M' must not create any cycles, so M' must not have any cycles.

Thus, M' is a subgraph of G with vertex set V that contains no cycles, so $M' \in \Sigma$. Also, since M' was constructed by adding a single edge to M, we know M is a subgraph of M'. That is,

$$M \leq M'$$

By definition of M as a maximal element of Σ, this implies

$$M = M'$$

However, $E_M \neq E_M \cup e_i = E_{M'}$ since $e_i \notin E_M$, so

$$M \neq M'$$

Thus, by assuming M is not connected, we have derived a contradiction, which completes the proof that M is connected.

Since M is connected, has vertex set V, and has no cycles, M is a spanning tree for $G = (V, E)$. This completes the proof that, for all connected graphs $G = (V, E)$, there exists a tree M that spans G. Since G_P is connected, this implies there exists a tree M_P which spans G_P. For any two vertices $v_1, v_2 \in V$, a spanning tree connects v_1 to v_2 with exactly one path (of finitely many edges). Thus, the existence of the spanning tree M_P for the graph G_P completes the proof that there is a way to connect all the planets using some of the roads such that there is exactly one path (finitely many roads) between any two planets.

Problem 5

(Bonus, 5 points). In this exercise, a close box in $\mathbb{R}^k$ is a product of k closed intervals

$$[a_1, b_1] \times \cdots \times [a_k, b_k]$$

Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is a function such that for any closed box $B \subseteq \mathbb{R}^m$, either $f^{-1}(B)$ or $\mathbb{R}^n \setminus f^{-1}(B)$ is countable (we could have two closed boxes $B_1, B_2 \subseteq \mathbb{R}^m$ such that $f^{-1}(B_1)$ is countable while $f^{-1}(B_2)$ is uncountable). Prove that there is a point $q \in \mathbb{R}^m$ such that $\mathbb{R}^n \setminus f^{-1}(q)$ is countable, meaning that f is almost everywhere equal to q in this cardinality sense.

Solution

For all $x = (x_1, ..., x_m) \in \mathbb{R}^m$, we have

$$x \in \big[\lfloor x_1 \rfloor, \lfloor x_1 \rfloor + 1\big] \times \cdots \times \big[\lfloor x_m \rfloor, \lfloor x_m \rfloor + 1\big]$$

and $\lfloor x_i \rfloor, \lfloor x_i \rfloor + 1 \in \mathbb{Z}$ for all $i \in \{1, ..., m\}$. Thus, we can cover all of $\mathbb{R}^m$ with closed cubes of side length 1, all of whom have vertices with exclusively integer coordinates. That is,

$$\mathbb{R}^m = \bigcup_{z=(z_1,...,z_m)\in\mathbb{Z}^m} [z_1 - 1, z_1] \times \cdots \times [z_m - 1, z_m]$$

Let $h_1 : \mathbb{N} \to \mathbb{Z}$ and $h_2 : \mathbb{Z} \to \mathbb{N}$ defined by

$$h_1(n) = \begin{cases} \frac{n}{2} \text{ if } n \equiv 0 \pmod 2 \\ -(\frac{n-1}{2}) \text{ if } n \equiv 1 \pmod 2 \end{cases} \qquad h_2(z) = \begin{cases} 2z \text{ if } z > 0 \\ 1 - 2z \text{ if } z \leq 0 \end{cases}$$

Note that

$$h_1 \circ h_2(z) = \begin{cases} h_1(2z) \text{ if } z > 0 \\ h_1(1 - 2z) \text{ if } z \leq 0 \end{cases} = \begin{cases} z \text{ if } z > 0 \\ -(1 - 2z - 1)/2 = z \text{ if } z \leq 0 \end{cases} = z$$

for all $z \in \mathbb{Z}$ (since $2z \equiv 0 \pmod 2$ and $1 - 2z \equiv 1 \pmod 2$ for all $z \in \mathbb{Z}$). Similarly,

$$h_2 \circ h_1(n) = \begin{cases} h_2(\frac{n}{2}) \text{ if } n \equiv 0 \pmod 2 \\ h_2(-(\frac{n-1}{2})) \text{ if } n \equiv 1 \pmod 2 \end{cases} = \begin{cases} n \text{ if } n \equiv 0 \pmod 2 \\ 1 + 2\frac{n-1}{2} = n \text{ if } n \equiv 1 \pmod 2 \end{cases} = n$$

for all $n \in \mathbb{N}$ (since $\frac{n}{2} > 0$ for all $n \in \mathbb{N}$ and $-\frac{n-1}{2} \leq 0$ for all $n \in \mathbb{N}$ (since $\frac{n-1}{2} \geq 0$ for all $n \in \mathbb{N}$)). Thus, we have

$$h_1 \circ h_2 = id_{\mathbb{Z}} \qquad h_2 \circ h_1 = id_{\mathbb{N}}$$

so we know from lecture that a h_1 is a bijection from $\mathbb{N}$ to $\mathbb{Z}$, and $|\mathbb{Z}| = \mathbb{N}|$. By definition, this implies $\mathbb{Z}$ is countable. By Proposition 1.7 in the notes, since

$$\mathbb{Z}^m = \underbrace{\mathbb{Z} \times \cdots \times \mathbb{Z}}_{m \text{ } \mathbb{Z}\text{'s}}$$

is a finite product of countable sets, Z^m is also countable. Thus, $\mathbb{R}^m$ is a countable union of closed cubes of side length 1.

From lecture, we know that $\mathbb{R}$ is uncountable. Note that $g : \mathbb{R} \to \mathbb{R}^n$ defined by $g(x) = \underbrace{(x, \cdots, x)}_{n\ x\text{'s}}$ is injective because

$$(x, \cdots, x) = (y, \cdots, y) \implies x = y$$

This implies $|R| \leq |R^n|$, so R^n is uncountable.
Since $f : \mathbb{R}^n \to \mathbb{R}^m$, we know $f(x) \in \mathbb{R}^m$ for all $x \in \mathbb{R}^n$. Thus,

$$f^{-1}(\mathbb{R}^m) = \mathbb{R}^n$$

by definition of the pre-image. Note also that

$$\begin{aligned} \mathbb{R}^n = f^{-1}(\mathbb{R}^m) = & \quad f^{-1}\Big(\bigcup_{z=(z_1,\ldots,z_m)\in\mathbb{Z}^m} [z_1 - 1, z_1] \times \cdots \times [z_m - 1, z_m]\Big) \\ = & \bigcup_{z=(z_1,\ldots,z_m)\in\mathbb{Z}^m} f^{-1}([z_1 - 1, z_1] \times \cdots \times [z_m - 1, z_m]) \end{aligned}$$

Note that the second equality follows by part (ii) of **Problem 1**. Thus, if

$$f^{-1}([z_1 - 1, z_1] \times \cdots \times [z_m - 1, z_m])$$

is countable for all $z \in \mathbb{Z}^m$, then $\mathbb{R}^n$ is a countable union of countable sets, so $\mathbb{R}^n$ must be countable by Proposition 1.5 from the notes. However, we already proved $\mathbb{R}^n$ is uncountable, so this would yield a contradiction. This means there exists some $y = (y_1, ..., y_m) \in \mathbb{Z}^m$ such that

$$f^{-1}([y_1 - 1, y_1] \times \cdots \times [y_m - 1, y_m])$$

is uncountable.
Consider the metric space $(\mathbb{R}^m, d)$, where d is the Euclidean distance metric defined by

$$d(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

for any two points $x = (x_1, \cdots, x_m), y = (y_1, \cdots, y_m) \in \mathbb{R}^m$. Note that

$$Y := [y_1 - 1, y_1] \times \cdots \times [y_m - 1, y_m]$$

is a closed m-dimensional cube with diameter

$$D_Y := diam(Y) := \sup_{x,y\in Y} d(x, y) = \sqrt{\underbrace{1^2 + \cdots + 1^2}_{m\ 1^2\text{'s}}} = \sqrt{m}$$

such that $f^{-1}(Y)$ is uncountable.
We claim that, from any closed m-dimensional cube B with diameter D_B such that $f^{-1}(B)$ is uncountable, we can find a nested, closed, m-dimensional cube B' of diameter $D_{B'} = \frac{D_B}{2}$ such that $f^{-1}(B')$ is uncountable.
Note that any m-dimensional cube

$$B = [b_{11}, b_{12}] \times \cdots \times [b_{m1}, b_{m2}]$$

of side length s can be split into 2^m smaller cubes with side length $\frac{s}{2}$ such that the union of the cubes is B and the interiors of the cubes are mutually disjoint. That is, if we define

$$A := \Big\{ [a_{11}, a_{12}] \times \cdots \times [a_{m1}, a_{m2}] \Big| (a_{i1}, a_{i2}) \in \Big\{ (b_{i1}, \frac{b_{i2} + b_{i1}}{2}), (\frac{b_{i2} + b_{i1}}{2}, b_{i2}) \Big\} \text{ for all } i \in \{1, ..., m\} \Big\}$$

and write

$$A = \{A_1, \cdots, A_{2^m}\}$$

since we know $|A| = 2^m$, then we can write

$$B = \bigcup_{i=1}^{2^m} A_i$$

If $f^{-1}A_i$ is countable for all $i \in \{1, ..., 2^m\}$, then

$$f^{-1}(B) = f^{-1}(\bigcup_{i=1}^{2^m} A_i) = \bigcup_{i=1}^{2^m} f^{-1}(A_i)$$

is a finite (and thus countable) union of finite sets, so $f^{-1}(B)$ must be countable by Proposition 1.5 from the notes (note that the second equality again follows by part (ii) of **Problem 1**). However, we assumed that $f^{-1}(B)$ is uncountable, so this yields a contradiction, and we conclude that $f^{-1}(A_i)$ must be *uncountable* for some $i^* \in \{1, ..., 2^m\}$.
Since, the side length of B is

$$s = b_{i2} - b_{i1} \text{ for any } i \in \{1, ..., m\}$$

we know B has diameter

$$D_B := \sup_{x,y \in B} d(x, y) = \sqrt{\underbrace{s^2 + \cdots + s^2}_{m \ s^2\text{'s}}} = \sqrt{ms^2} = s\sqrt{m} = (b_{i2} - b_{i1})\sqrt{m}$$

We can easily compute that A_{i^*} has side length

$$\frac{b_{i2} + b_{i1}}{2} - b_{i1} = \frac{b_{i2} - b_{i1}}{2} = \frac{s}{2} = b_{i2} - \frac{b_{i2} + b_{i1}}{2}$$

which implies A_{i^*} has diameter

$$D_{A_{i^*}} := \sup_{x,y \in A_{i^*}} d(x, y) = \sqrt{\underbrace{\frac{s^2}{4} + \cdots + \frac{s^2}{4}}_{m \ \frac{s^2}{4}\text{'s}}} = \sqrt{m\frac{s^2}{4}} = \frac{s}{2}\sqrt{m} = \frac{1}{2}(b_{i2} - b_{i1})\sqrt{m} = \frac{D_B}{2}$$

Letting $B' = A_{i^*} \subseteq B$ completes the proof that, from any closed m-dimensional cube B of diameter D_B such that $f^{-1}(B)$ is uncountable, we can construct a nested, closed, m-dimensional cube B' of diameter $D_{B'} = \frac{D_B}{2}$ such that $f^{-1}(B')$ is uncountable.
By induction, since Y is a closed m-dimensional cube of diameter $D_Y = \sqrt{M}$ such that $f^{-1}(Y)$ is uncountable, we can create the following (countably) infinite sequence $B_1, B_2, B_3, ...$, defined by

1. $B_1 = Y = [b_{11}^1, b_{12}^1] \times \cdots \times [b_{m1}^1, b_{m2}^1]$
2. For all $n \in \mathbb{N}$,

$$B_{n+1} = [b_{11}^{n+1}, b_{12}^{n+1}] \times \cdots \times [b_{m1}^{n+1}, b_{m2}^{n+1}]$$

such that

$$(b_{i1}^{n+1}, b_{i2}^{n+1}) \in \left\{ \left(b_{i1}^n, \frac{b_{i2}^n + b_{i1}^n}{2}\right), \left(\frac{b_{i2}^n + b_{i1}^n}{2}, b_{i2}^n\right) \right\}$$

for all $i \in \{1, ..., m\}$ and $f^{-1}(B_{n+1})$ is uncountable.

so B_n is a closed m-dimensional cube such that $f^{-1}(B_n)$ is uncountable for all $n \in \mathbb{N}$.
By induction, since $D_{B_{n+1}} := diam(B_{n+1}) = \frac{diam(B_n)}{2} =: \frac{D_{B_n}}{2}$ and $D_{B_1} = D_Y = \sqrt{m}$, we have

$$D_{B_n} := diam(B_n) = \frac{\sqrt{m}}{2^{n-1}}$$

The numerator is constant while the denominator tends to infinity as $n \to \infty$, so we can write

$$\lim_{n\to\infty} diam(B_n) = \lim_{n\to\infty} \sup_{x,y\in B_n} d(x,y) = \lim_{n\to\infty} \frac{\sqrt{m}}{2^{n-1}} = 0$$

This directly implies

$$d(x,y) = 0 \text{ for all } x, y \in \lim_{n\to\infty} B_n$$

By definition of a metric, $d(x,y) = 0 \iff x = y$. Thus, we have

$$x = y \text{ for all } x, y \in \lim_{n\to\infty} B_n$$

Only the empty set $\emptyset$ and singleton sets of the form $\{q\}$ could satisfy this statement for an arbitrary set. Since B_n is a closed m-dimensional cube by definition, it cannot be the emptyset, so we must have

$$\lim_{n\to\infty} B_n = \{q\} \text{ for some } q \in \mathbb{R}^m$$

Also, since $f^{-1}(\lim_{n\to\infty} B_n)$ is uncountable by definition, we know $\mathbb{R}^n \setminus f^{-1}(\lim_{n\to\infty} B_n)$ is countable (by the assumption on f from the problem statement). Since $\lim_{n\to\infty} B_n = \{q\}$ for some $q \in \mathbb{R}^m$, this completes the proof that there exists a point $q \in \mathbb{R}^m$ such that $\mathbb{R}^n \setminus f^{-1}(q)$ is countable.

Assignment 3

Problem 1

(5 points). Let S be the set of functions $N \to \{0,1\}$ such that for every $f \in S$, both $f^{-1}(0)$ and $f^{-1}(1)$ are countably infinite. Prove that S is uncountable.

Hint: When is $f \notin S$?

Solution

First, note that $S \subseteq \{0,1\}^{\mathbb{N}}$ by definition.
Claim: $\{0,1\}^{\mathbb{N}}$ is uncountable.

Proof. For all $A \in \mathcal{P}(\mathbb{N})$ (i.e. for all $A \subseteq \mathbb{N}$), define $f_A : \mathbb{N} \to \{0,1\}$ such that

$$f_A(n) = \begin{cases} 1 \text{ if } n \in A \\ 0 \text{ if } n \notin A \end{cases}$$

for all $n \in \mathbb{N}$.
Now, consider $g : \mathcal{P}(\mathbb{N}) \to \{0,1\}^{\mathbb{N}}$ defined by

$$g(A) = f_A$$

for all $A \in \mathcal{P}(\mathbb{N})$. Then, for all $A_1, A_2 \in \mathcal{P}(\mathbb{N})$,

$$g(A_1) = g(A_2) \implies f_{A_1} = f_{A_2} \implies f_{A_1}(n) = f_{A_2}(n)$$

for all $n \in \mathbb{N}$. Thus, if $g(A_1) = g(A_2)$, then for all $n \in \mathbb{N}$, we have

$$n \in A_1 \implies f_{A_1}(n) = 1 \implies f_{A_2}(n) = 1 \implies n \in A_2$$

Similarly, for all $n \in \mathbb{N}$, we have

$$n \in A_2 \implies f_{A_2}(n) = 1 \implies f_{A_1}(n) = 1 \implies n \in A_2$$

Thus,

$$(g(A_1) = g(A_2)) \implies \forall n \in \mathbb{N}((n \in A_1) \iff (n \in A_2))$$

By the axiom of extensionality,

$$\forall n \in \mathbb{N}((n \in A_1) \iff (n \in A_2)) \implies (A_1 = A_2)$$

so we have

$$(g(A_1) = g(A_2)) \implies A_1 = A_2$$

That is, g is an injection from $\mathcal{P}(\mathbb{N}) \to \{0,1\}^{\mathbb{N}}$, so

$$|\mathcal{P}(\mathbb{N})| \leq |\{0,1\}^{\mathbb{N}}|$$

By Cantor's Theorem (Theorem 1.1 in the notes),

$$|\mathbb{N}| < |\mathcal{P}(\mathbb{N})|$$

Combining these two inequalities, we have

$$|\mathbb{N}| < |\mathcal{P}(\mathbb{N})| \leq |\{0,1\}^{\mathbb{N}}| \implies |\mathbb{N}| < |\{0,1\}^{\mathbb{N}}|$$

By the definition of countability, this completes the proof that $\{0,1\}^{\mathbb{N}}$ is uncountable. □

Note that

$$S = \{f \in \{0,1\}^{\mathbb{N}} | f^{-1}(0) \text{ and } f^{-1}(1) \text{ are both countably infinite}\}$$

and define

$$S' := \{0,1\}^{\mathbb{N}} \setminus S$$

so that

$$\{0,1\}^{\mathbb{N}} = S' \cup S$$

If a set is not countably infinite, it is either finite or uncountably infinite. Thus,

$$S' = \{f \in \{0,1\}^{\mathbb{N}} | f^{-1}(0) \text{ or } f^{-1}(1) \text{ is finite or uncountably infinite}\}$$

For all $f \in \{0,1\}^{\mathbb{N}}$ (and thus all $f \in S$), we know

$$f^{-1}(0) \cup f^{-1}(1) = f^{-1}(\{0,1\}) = \mathbb{N}$$

By definition of countability, we know $\mathbb{N}$ is countably infinite. If $f^{-1}(0)$ is uncountable, then

$$f^{-1}(0) \cup f^{-1}(1) = f^{-1}(\{0,1\}) = \mathbb{N}$$

is uncountable, a direct contradiction of the definition of countability. Similarly, if $f^{-1}(1)$ is uncountable, then

$$f^{-1}(1) \cup f^{-1}(0) = f^{-1}(\{0,1\}) = \mathbb{N}$$

is uncountable, another contradiction. Thus, we can rewrite the compliment of S as

$$S' = \{f \in \{0,1\}^{\mathbb{N}} | f^{-1}(0) \text{ or } f^{-1}(1) \text{ is finite}\}$$

If both $f^{-1}(1)$ and $f^{-1}(0)$ are finite, then

$$f^{-1}(1) \cup f^{-1}(0) = f^{-1}(\{0,1\}) = \mathbb{N}$$

is finite (as the finite union of finite sets). Thus, for all $f \in \{0,1\}^{\mathbb{N}}$,

$$(f^{-1}(0) \text{ is finite }) \implies (f^{-1}(1) \text{ is countably infinite })$$

and

$$(f^{-1}(1) \text{ is finite }) \implies (f^{-1}(0) \text{ is countably infinite })$$

This means we can rewrite S' as

$$S' = \{f \in \{0,1\}^{\mathbb{N}} | \text{ exactly one of } f^{-1}(0) \text{ and } f^{-1}(1) \text{ is finite}\}$$

For all $f \in \{0,1\}^{\mathbb{N}}$, if $f^{-1}(0)$ is finite, then

$$f(n) = \begin{cases} 0 \text{ if } n \in A \\ 1 \text{ if } n \notin A \end{cases}$$

for some $A \subseteq \mathbb{N}$ such that $A = \{a_1, ..., a_k\}$ is finite. Similarly, if $f^{-1}(1)$ is finite, we have

$$f(n) = \begin{cases} 1 \text{ if } n \in A \\ 0 \text{ if } n \notin A \end{cases}$$

for some $A \subseteq \mathbb{N}$ such that $A = \{a_1, ..., a_k\}$ is finite. Thus, for all finite subsets $A \subseteq \mathbb{N}$, we can define f_0^A and f_1^A, both mapping from $\mathbb{N}$ to $\{0,1\}$, such that

$$f_0^A(n) = \begin{cases} 0 \text{ if } n \in A \\ 1 \text{ if } n \notin A \end{cases} \qquad \forall n \in \mathbb{N}$$

and

$$f_1^A(n) = \begin{cases} 1 \text{ if } n \in A \\ 0 \text{ if } n \notin A \end{cases} \qquad \forall n \in \mathbb{N}$$

Then we can rewrite S' once more as

$$S' = \{f_0^A, f_1^A \in \{0,1\}^{\mathbb{N}} | A \subseteq \mathbb{N} \text{ is finite}\} = \{f_i^A \in \{0,1\}^{\mathbb{N}} | i \in \{0,1\}, A \subseteq \mathbb{N} \text{ is finite}\}$$

Claim: S', as defined above, is countable.

Proof. First, note that for any $n, m \in \mathbb{N}$,

$$n \neq m \implies (n < m) \vee (n > m)$$

Thus, for any arbitrary finite subset $A \subseteq \mathbb{N}$ of size $k \in \mathbb{N}$, we can number the (necessarily unique by the definition of a set) k elements $a_1, ..., a_k$ of A such that

$$a_1 < \cdots < a_k$$

That is, we can write

$$A = \{a_1, \cdots, a_k | a_1 < \cdots < a_k, \quad k \in \mathbb{N}\}$$

for all such finite subsets $A \subseteq \mathbb{N}$. Consider $h : S' \to \mathbb{Z}$ such that, for all $f_i^A \in S'$,

$$h(f_i^A) = h(f_i^{\{a_1, \cdots, a_k | a_1 < \cdots < a_k, k \in \mathbb{N}\}}) = (-1)^i p_1^{a_1} \cdots p_k^{a_k}$$

where p_i is the i'th smallest prime (i.e. $p_1, p_2, p_3, \ldots = 2, 3, 5, \ldots$). Note this implies

$$h(f_0^\emptyset) = (-1)^0 = 1, \qquad h(f_1^\emptyset) = (-1)^1 = -1, \text{ and} \qquad |h(f_i^A)| \geq 2$$

for all $i \in \{0,1\}$ and nonempty finite subsets $A \subseteq \mathbb{N}$.
Thus,

$$h(f_i^A) = -1 = h(f_j^B) \iff (i = j = 1) \wedge (A = B = \emptyset)$$

and

$$h(f_i^A) = 1 = h(f_j^B) \iff (i = j = 0) \wedge (A = B = \emptyset)$$

Also, without loss of generality, if $A = \emptyset$ and $B \neq \emptyset$, then

$$h(f_i^A) \neq h(f_j^B)$$

since $|h(f_i^A)| = 1 < 2 \leq |h(f_j^B)|$. Thus, for all $i, j \in \{0,1\}$ and all finite subsets $A, B \subseteq \mathbb{N}$ such that $(A = \emptyset) \vee (B = \emptyset)$,

$$h(f_i^A) = h(f_j^B) \implies (i = j) \wedge (A = B)$$

For any $i, j \in \{0,1\}$ and any *nonempty* finite subsets $A, B \subseteq \mathbb{N}$,

$$(h(f_i^A) = h(f_j^B)) \implies ((-1)^i p_1^{a_1} \cdots p_k^{a_k} = (-1)^j p_1^{b_1} \cdots p_l^{b_l})$$

Note that $p_i > 0$ for all $i \in \mathbb{N}$, so $p_i^n > 0$ for all $i, n \in \mathbb{N}$, so $p_1^{n_1} \cdots p_k^{n_k} > 0$ for all $k, n_1, ..., n_k \in \mathbb{N}$. Thus, if $i = 0$ and $j = 1$,

$$0 < p_1^{a_1} \cdots p_k^{a_k} = (-1)^i p_1^{a_1} \cdots p_k^{a_k} = (-1)^j p_1^{b_1} \cdots p_l^{b_l} = -p_1^{b_1} \cdots p_l^{b_l} < 0$$

which contradicts the law of excluded middle since $0 < 0 \implies (0 \leq 0) \wedge (0 \neq 0)$. Similarly, if $i = 1$ and $j = 0$, then

$$0 > -p_1^{a_1} \cdots p_k^{a_k} = (-1)^i p_1^{a_1} \cdots p_k^{a_k} = (-1)^j p_1^{b_1} \cdots p_l^{b_l} = p_1^{b_1} \cdots p_l^{b_l} > 0$$

so we reach the same contradiction. Thus,

$$h(f_i^A) = h(f_j^B) \implies (-1)^i p_1^{a_1} \cdots p_k^{a_k} = (-1)^j p_1^{b_1} \cdots p_l^{b_l} \implies i = j \implies p_1^{a_1} \cdots p_k^{a_k} = p_1^{b_1} \cdots p_l^{b_l}$$

with the last implication following since $(-1)^i \neq 0$ for all $i \in \{0, 1\}$. Since $a_1, ..., a_k, b_1, ..., b_l \in \mathbb{N}$, $p_1^{a_1} \cdots p_k^{a_k}$ has exactly a_x factors of p_x for all $x \in \{1, ..., k\}$ and $p_1^{b_1} \cdots p_l^{b_l}$ has at exactly b_y factors of p_y for all $y \in \{1, ..., l\}$. By the fundamental theorem of arithmetic, each natural number $n \geq 2$ has a unique prime factorization. Thus, for

$$p_1^{a_1} \cdots p_k^{a_k} = p_1^{b_1} \cdots p_l^{b_l}$$

to hold, we must have $k = l$ and $a_x = b_x$ for all $x \in \{1, ..., k\} = \{1, ..., l\}$. That is, for all $f_i^A, f_j^B \in S'$ such that $A, B \subseteq \mathbb{N}$ are nonempty finite subsets,

$$(h(f_i^A) = h(f_j^B)) \implies (i = j) \wedge (A := \{a_1, ..., a_k\} = \{b_1, ..., b_l\} =: B)$$

We already show that

$$(h(f_i^A) = h(f_j^B)) \implies ((i = j) \wedge (A = B))$$

for all $i, j \in \{0, 1\}$ and $A, B \subseteq \mathbb{N}$ such that $(A = \emptyset) \wedge (B = \emptyset)$. Thus, we have

$$(h(f_i^A) = h(f_j^B)) \implies ((i = j) \wedge (A = B)) \implies (f_i^A = f_j^B)$$

for all $i, j \in \{0, 1\}$ and all finite subsets (empty or nonempty) $A, B \subseteq \mathbb{N}$. That is, h is an injection from S' to $\mathbb{Z}$, so

$$|S'| \leq |\mathbb{Z}|$$

Now, consider $H : \mathbb{Z} \to \mathbb{N}$ defined by

$$H(z) = \begin{cases} 1 - 2z \text{ if } z \leq 0 \\ 2z \text{ if } z > 0 \end{cases} \qquad \forall z \in \mathbb{Z}$$

Note that, for all $z \in \mathbb{Z}$,

$$1 - 2z \equiv 1 - 2z + 2z \equiv 1 \pmod 2$$

while

$$2z \equiv 2z - 2z \equiv 0 \pmod 2$$

so

$$(H(z_1) = H(z_2)) \implies ((z_1, z_2 \leq 0) \vee (z_1, z_2 > 0))$$

This combines with the definition of H to imply

$$\begin{aligned}
(H(z_1) = H(z_2)) \implies & \begin{cases} 1 - 2z_1 = 1 - 2z_2 \text{ if } z_1, z_2 \leq 0 \\ 2z_1 = 2z_2 \text{ if } z_1, z_2 > 0 \end{cases} \\
\implies & \begin{cases} 1 - 2z_1 - 1 = 1 - 2z_2 - 1 \text{ if } z_1, z_2 \leq 0 \\ 2^{-1} \cdot 2z_1 = 2^{-1} \cdot 2z_2 \text{ if } z_1, z_2 > 0 \end{cases} \\
\implies & \begin{cases} -2z_1 = -2z_2 \text{ if } z_1, z_2 \leq 0 \\ z_1 = z_2 \text{ if } z_1, z_2 > 0 \end{cases} \\
\implies & \begin{cases} (-2)^{-1} \cdot (-2z_1) = (-2)^{-1} \cdot (-2z_2) \text{ if } z_1, z_2 \leq 0 \\ z_1 = z_2 \text{ if } z_1, z_2 > 0 \end{cases} \\
\implies & \begin{cases} z_1 = z_2 \text{ if } z_1, z_2 \leq 0 \\ z_1 = z_2 \text{ if } z_1, z_2 > 0 \end{cases} \\
\implies & z_1 = z_2
\end{aligned}$$

That is H is an injection from Z to $\mathbb{N}$, so

$$|\mathbb{Z}| \leq |\mathbb{N}|$$

Putting our inequalities together yields

$$|S'| \leq |\mathbb{Z}| \leq |\mathbb{N}| \implies |S'| \leq |\mathbb{N}|$$

By the definition of countability, this completes the proof that $S' := \{0,1\}^{\mathbb{N}} \setminus S$ is countable. □

Now, we finally claim that S is uncountable.

Proof. Assume to the contrary that S is countable. Recall that

$$\{0,1\}^{\mathbb{N}} = S \cup (\{0,1\}^{\mathbb{N}} \setminus S) =: S \cup S'$$

and we already showed $\{0,1\}^{\mathbb{N}}$ is uncountable. Since we just proved S' is countable, our assumption that S is also countable implies

$$S \cup S' = \{0,1\}^{\mathbb{N}}$$

is a finite (and thus countable) union of countable sets. By Proposition 1.5 in the notes, this implies $\{0,1\}^{\mathbb{N}}$ is countable. Since $\{0,1\}^{\mathbb{N}}$ is uncountable, this contradicts the law of excluded middle, so we know S cannot be countable. The conclusion that S must be uncountable follows by contradiction and the definition of uncountability. □

Problem 2

(5 points). Recall that a $\mathbb{Q}$-vector space V is a set (whose elements are called vectors) with an addition operation $+ : V \times V \to V$ along with a scaling operation $\mathbb{Q} \times V \to V$ satisfying all properties listed below. Given $\lambda \in \mathbb{Q}$, $v \in V$, we denote scalar multiplication $\lambda \cdot v$ or λv such that

($+_1$) There is a zero vector $0 \in V : \forall v \in V$, $0 + v = v + 0 = v$,

($+_2$) $+$ is associative: $\forall u, v, w \in V$, $u + (v + w) = (u + v) + w$,

($+_3$) additive inverses exist: $\forall v \in V$, $\exists w \in V$ such that $v + w = w + v = 0$,

($+_4$) $+$ is commutative: $\forall u, v \in V$, $u + v = v + u$,

and

($\cdot_1$) $\forall \lambda \in \mathbb{Q}$, $\forall u, v \in V$, $\lambda(u + v) = \lambda u + \lambda v$,

($\cdot_2$) $\forall \lambda, \mu \in \mathbb{Q}$, $\forall v \in V$, $(\lambda + \mu)v = \lambda v + \mu v$,

($\cdot_3$) $\forall \lambda, \mu \in \mathbb{Q}$, $\forall v \in V$, $\lambda(\mu v) = (\lambda\mu)v$,

($\cdot_4$) $\forall v \in V$, $1 \cdot v = v$, that is, 1 acts trivially on v.

The first three $+$ axioms say that $(V, +)$ is a group. Adding the last axiom says that $(V, +)$ is an abelian group. Prove the following:

(a) in every $\mathbb{Q}$-vector space V, zero vectors are unique, that is, if $0_1, 0_2 \in V$, $0_1 = 0_2$,

(b) additive inverses are unique, that is, for any $v \in V$, if $w_1, w_2 \in V$ are additive inverses to v, then $w_1 = w_2$,

(c) for every $v \in V$, let $-v := (-1) \cdot v$. Prove that $-v$ is the additive inverse to v,

(d) if $\lambda \in \mathbb{Q}^x := \mathbb{Q} \setminus \{0\}$ such that $\lambda v = 0$, then $v = 0$,

(e) if $v \in V$, then $0 \cdot v = 0$, where 0 on the left is in $\mathbb{Q}$, 0 on the right is the zero vector of V.

Solution

(a) Let V be an arbitrary $\mathbb{Q}$-vector space and consider two arbitrary 0 vectors $0_1, 0_2 \in V$. By $(+_1)$, since $v_1 = v_1 + 0$ and $0 + v_2 = v_2$ for all $v_1, v_2 \in V$, we have

$$0_1 \underbrace{=}_{v_1 = v_1 + 0} 0_1 + 0_2 \underbrace{=}_{0 + v_2 = v_2} 0_2$$

Thus, in any $\mathbb{Q}$-vector space V, if two 0 vectors $0_1, 0_2 \in V$, we must have $0_1 = 0_2$. This completes the proof that zero vectors are unique in every $\mathbb{Q}$-vector space.

(b) Let V be an arbitrary $\mathbb{Q}$-vector space and $v \in V$ an arbitrary vector therein. Suppose there exist two additive inverses $w_1, w_2 \in V$ to v. That is, $w_1 + v = 0 = v + w_2$. Then by $(+_1)$, $(+_2)$, and $(+_3)$,

$$w_1 \underbrace{=}_{v = v + 0} w_1 + 0 \underbrace{=}_{v + w_2 = 0} w_1 + (v + w_2) \underbrace{=}_{u + (v + w) = (u + v) + w} (w_1 + v) + w_2 \underbrace{=}_{w_1 + v = 0} 0 + w_2 \underbrace{=}_{0 + v = v} w_2$$

So if two additive inverses w_1, w_2 exist to any vector v in any $\mathbb{Q}$-vector space V, we must have $w_1 = w_2$. This completes the proof that additive inverses are unique in every $\mathbb{Q}$-vector space.

(c) Let V be an arbitrary $\mathbb{Q}$-vector space and $v \in V$ an arbitrary vector therein, as before. By $(\cdot_2)$ and $(\cdot_4)$,

$$v + -v := v + (-1) \cdot v \underbrace{=}_{v = 1 \cdot v} 1 \cdot v + (-1) \cdot v \underbrace{=}_{\lambda v + \mu v = (\lambda + \mu) v} (1 + -1) \cdot v$$

In Lemma 2.9 of the lecture notes, we defined -1 to be the additive inverse of 1 in $\mathbb{R}$, and we said this holds true for $\mathbb{Q}$ too. That is, $1 + -1 = 0$, so we have

$$v + -v = 0 \cdot v$$

By part (e), we know $0 \cdot v = 0$ for all $v \in V$, which implies

$$v + -v = 0$$

Note: We have not yet proven part (e), but its proof will not depend on part (c), so we can use its result to establish part (c).

Since we know additive inverses are unique by part (b), and $-v$ satisfies the definition of the additive inverse to v from $(+_3)$ since $v + -v = 0$, this completes the proof that $-v := -1 \cdot v$ is the additive inverse to v for all $v \in V$.

(d) From the lecture notes, we know that, for any $r \in \mathbb{R}^x := \mathbb{R} \setminus \{0\}$, there exists a multiplicative inverse $r^{-1} \in \mathbb{R}^x$ such that $r \cdot r^{-1} = r^{-1} \cdot r = 1$, and we know that the same holds for $\mathbb{Q}$. Thus, for any $\lambda \in \mathbb{Q}^x := \mathbb{Q} \setminus \{0\}$, $\exists \lambda^{-1} \in \mathbb{Q}^x$ such that $\lambda \cdot \lambda^{-1} = \lambda^{-1} \cdot \lambda = 1$. Applying $(\cdot_3)$ and $(\cdot_4)$, if $\lambda v = 0$, we find

$$v \underbrace{=}_{(\cdot_4)} 1 \cdot v \underbrace{=}_{\lambda \lambda^{-1} = 1} (\lambda^{-1} \lambda) v \underbrace{=}_{(\cdot_3)} \lambda^{-1} (\lambda v) \underbrace{=}_{\lambda v = 0} \lambda^{-1} \cdot 0$$

Claim: For all $q \in Q^x$ and all $\mathbb{Q}$-vector spaces V with zero vector 0, $q \cdot 0 = 0$.

Proof. We have

$$q \cdot 0 \underbrace{=}_{(+_1)} q \cdot (0 + 0) \underbrace{=}_{(\cdot_1)} q \cdot 0 + q \cdot 0$$

Since $q \cdot 0 \in V$, $(+_3)$ implies the existence of some $w \in V$ such that $(q \cdot 0) + w = 0 \in V$. Adding this to both sides of our equation yields

$$0 \underbrace{=}_{(q \cdot 0) + w = 0} (q \cdot 0) + w \underbrace{=}_{q \cdot 0 = q \cdot 0 + q \cdot 0} (q \cdot 0 + q \cdot 0) + w \underbrace{=}_{(+_2)} q \cdot 0 + (q \cdot 0 + w) \underbrace{=}_{q \cdot 0 + w = 0} q \cdot 0$$

This completes the proof that $q \cdot 0 = 0$ for all $q \in \mathbb{Q}^x$ and all $\mathbb{Q}$-vector spaces V with zero vector 0. $\square$

Since $\lambda^{-1} \in \mathbb{Q}^x$, we find that $\lambda v = 0$, $\lambda \in \mathbb{Q}^x$ implies

$$v = \lambda^{-1} \cdot 0 = 0$$

This completes the proof that, if $\lambda \in \mathbb{Q}^x$ such that $\lambda v = 0$, then $v = 0$.

(e) Consider an arbitrary $\mathbb{Q}$-vector space V and an arbitrary vector $v \in V$ therein. From the lecture notes, we know that $0 + 0 = 0$ for $0 \in \mathbb{R}$ and that the same holds for $0 \in \mathbb{Q}$. Thus,

$$0 \cdot v \underbrace{=}_{0=0+0} (0+0) \cdot v \underbrace{=}_{(\cdot_2)} 0 \cdot v + 0 \cdot v$$

Since $0 \cdot v \in V$, we know there exists $w \in V$ such that $0 \cdot v + w = 0$. Adding w to both sides of our equation yields

$$0 \underbrace{=}_{0 \cdot v + w = 0} 0 \cdot v + w \underbrace{=}_{0 \cdot v = 0 \cdot v + 0 \cdot v} (0 \cdot v + 0 \cdot v) + w \underbrace{=}_{(+_2)} 0 \cdot v + (0 \cdot v + w) \underbrace{=}_{0 \cdot v + w = 0} 0 \cdot v$$

This completes the proof that, if $v \in V$, $0 \cdot v = 0$, where $0 \in \mathbb{Q}$ on the left and $0 \in V$ on the right.

Problem 3

(5 points). Suppose V is a $\mathbb{Q}$-vector space. A set of vectors $\{v_i | i \in I\} \subseteq V$ is said to be *Q-linearly dependent* if there are $\{\lambda_i, i \in I\} \subseteq \mathbb{Q}$, not all 0 and *all but finitely* many equal to zero, such that

$$\sum_{i \in I} \lambda_i v_i = 0.$$

In words, you can *non-trivially* express the zero vector as a linear combination of finitely many of the vectors. Otherwise, the set of vectors is said to be $\mathbb{Q}$-*linearly independent*. A $\mathbb{Q}$-linearly independent set of vectors $\mathcal{B} = \{b_\alpha | \alpha \in A\} \subseteq V$ is a *basis* if every $v \in V$ is a $\mathbb{Q}$-linear combination of finitely many of the elements of $\mathcal{B}$.

Note that this is a generalization of the usual definition of linear independence and basis for vector spaces over $\mathbb{R}$ that you have seen before.

Prove that every $\mathbb{Q}$-vector space has a basis using Zorn's lemma.

Hint: See notes.

Solution

Let

$$\Sigma = \{A \subseteq V | \text{ A is } \mathbb{Q}\text{-linearly independent}\}$$

be the set of all $\mathbb{Q}$-linearly independent subsets of V.
Claim: The pair $(\Sigma, \subseteq)$ forms a poset (partially ordered set).

Proof. We will show this pair satisfies reflexivity, anti-symmetry, and transitivity.

(i) *(Reflexivity).* For all $\mathbb{Q}$-linearly independent subsets $A \subseteq V$, we trivially have $A \subseteq A$. Thus, reflexivity holds under $\subseteq$ for all $A \in \Sigma$.

(ii) *(Anti-Symmetry).* For all $A, B \in \Sigma$,

$$(A \subseteq B) \wedge (B \subseteq A) \implies (\forall a \in A, a \in B) \wedge (\forall a \in B, a \in A) \implies \forall a((a \in A) \iff (a \in B)) \implies A = B$$

with the last implication following by the axiom of extensionality. Thus, anti-symmetry holds under $\subseteq$ for all $A, B \in \Sigma$.

(iii) *(Transitivity).* For all $A, B, C \in \Sigma$,

$$(A \subseteq B) \wedge (B \subseteq C) \implies ((\forall a \in A, a \in B) \wedge (\forall a \in B, a \in C)) \implies (\forall a \in A, a \in C) \implies A \subseteq C$$

so transitivity also holds under $\subseteq$ for all $A, B, C \in \Sigma$.

This completes the proof that $(\Sigma, \subseteq)$ is a poset. □

To apply Zorn's Lemma, we must first show that every chain C has an upper bound. Consider an arbitrary chain $C = \{A_i | i \in I\} \subseteq \Sigma$ and consider the set $A := \bigcup_{i \in I} A_i$. Note that, for all $i \in I$, by an inductive application of the definition of set union, we know

$$a \in A_i \implies a \in \bigcup_{i \in I} A_i =: A$$

Thus, for all $A_i \in C$, we have $A_i \subseteq A$. By Definition 1.27 from the notes, if $A \in \Sigma$, then A is an upper bound to C.
Claim: $A \in \Sigma$.

Proof. Since $C \subseteq \Sigma$, for all $A_i \in C$, $A_i \in \Sigma$. By definition of Σ, this means A_i is a $\mathbb{Q}$-linearly independent subset of V for all $A_i \in C$. By definition of the union of sets,

$$a \in \bigcup_{i \in I} A_i \implies a \in A_i \text{ for some } i \in I$$

Since $A_i \subseteq V$ for all $i \in I$, this means

$$a \in \bigcup_{i \in I} A_i \implies a \in V$$

Thus, $A := \bigcup_{i \in I} A_i$ is a subset of V. It remains to show A is also $\mathbb{Q}$-linearly independent. Assume to the contrary that A is $\mathbb{Q}$-linearly dependent. That is, there exists a finite subset $\{a_1, ..., a_k\} \subseteq A$ of $k \in \mathbb{N}$ vectors for which there exists k scalars $\lambda_1, .., \lambda_k \in \mathbb{Q}$ such that $\lambda_j \neq 0$ for all $j \in \{1, ..., k\}$ and

$$\sum_{j=1}^{k} \lambda_j a_j = 0$$

Since $a_j \in A := \bigcup_{i \in I} A_i$ for all $j \in \{1, ..., k\}$, we know that for each $a_j \in \{a_1, ..., a_k\}$, $a_j \in A_i$ for some $i \in I$. Thus, we can find k (not necessarily distinct) $A_{i_1}, ..., A_{i_k} \in C$ such that $a_j \in A_{i_j}$ for all $j \in \{1, ..., k\}$.
Claim: For all $k \in \mathbb{N}$ and for any collection of k (not necessarily distinct) $A_{i_1}, ..., A_{i_k} \in C$, there exists an A_{i_n} such that

$$A_{i_j} \subseteq A_{i_n} \text{ for all } j \in \{1, ..., k\}$$

Proof. We induct on k.
Base Case: $k = 1$, $A_1 \subseteq A_1$ by the reflexivity of the $\subseteq$ relation, so the claim holds for the base case.
Inductive Hypothesis: Assume that the claim holds for all $1 \leq k \leq n$.
Inductive Step: Consider $k = n + 1$. By the inductive hypothesis, we know there exists a $p \in \{1, ..., n\}$ such that

$$A_{i_j} \subseteq A_{i_p} \text{ for all } j \in \{1, ...n\}$$

Since C is a chain, and $A_{i_j} \in C$ for all $j \in \{1, ..., n+1\}$, we know

$$(A_{i_p} \subseteq A_{i_{n+1}}) \vee (A_{i_{n+1}} \subseteq A_{i_p})$$

If $A_{i_p} \subseteq A_{i_{n+1}}$, then

$$A_{i_j} \subseteq A_{i_{n+1}} \text{ for all } j \in \{1, ..., n+1\}$$

by the reflexivity and transitivity of the $\subseteq$ relation. On the other hand, if $A_{i_{n+1}} \subseteq A_{i_p}$, then

$$A_{i_j} \subseteq A_{i_p} \text{ for all } j \in \{1, ..., n+1\}$$

since we already know $A_{i_j} \subseteq A_{i_p}$ for all $j \in \{1, ..., n\}$ by the inductive hypothesis. Thus, in either case, we can always find a $l \in \{1, ..., n+1\}$ such that $A_{i_j} \subseteq A_{i_l}$ for all $j \in \{1, ..., n+1\}$. The conclusion that there exists such an A_{i_l} in any collection of k (not necessarily distinct) $A_{i_1}, ..., A_{i_k} \in C$ follows for all $k \in \mathbb{N}$ by induction. □

Recall that we found k not necessarily distinct $A_{i_1}, ..., A_{i_k} \in C$ such that $a_j \in A_{i_j}$ for all $j \in \{1, ..., k\}$. From the result we just proved, we know there exists some $A^* \in \{A_{i_1}, ..., A_{i_k}\}$ such that

$$A_{i_j} \subseteq A^* \text{ for all } j \in \{1, ..., k\}$$

Define

$$A^{(k)} = A^*$$

recursively define

$$A^{(x)} \text{ s.t. } A_{i_j} \subseteq A^{(x)} \text{ for all } A_{i_j} \in \{A_{i_1}, ..., A_{i_k}\} \setminus \{A^{(x+1)}, ..., A^{(k)}\}$$

The result we just proved implies we can do this for all $x \in \{1, ..., k-1\}$. After doing so, we get

$$\{A^{(1)}, ..., A^{(k)}\} = \{A_{i_1}, ..., A_{i_k}\}$$

and

$$A^{(1)} \subseteq \cdots \subseteq A^{(k)}$$

Recall that, for all $j \in \{1, ..., k\}$, $a_j \in A_{i_j}$. Since

$$\{A^{(1)}, ..., A^{(k)}\} = \{A_{i_1}, ..., A_{i_k}\}$$

this implies that, for all $j \in \{1, ..., k\}$, there exists some $x \in \{1, ..., k\}$ such that $a_j \in A^{(x)}$. Since

$$A^{(1)} \subseteq \cdots \subseteq A^{(k)}$$

$a_j \in A^{(x)} \implies A_j \in A^{(k)}$, so we know

$$a_j \in A^{(k)}$$

for all $j \in \{1, ..., k\}$. However, by assumption, this means there exists $\lambda_1, ..., \lambda_k \in \mathbb{Q}$ such that $\lambda_j \neq 0$ for all $j \in \{1, ..., k\}$ and

$$\sum_{j=1}^{k} \lambda_j a_j = 0$$

That is, the 0 vector can be produced as a *non-trivial* rational linear combination of vectors $a_1, ..., a_k \in A^{(k)}$. By definition, this implies $A^{(k)}$ is $\mathbb{Q}$-linearly dependent. However, recall that $A^{(k)} := A_{i_j} \in C$ for some $j \in \{1, ..., k\}$, and $C \subseteq \Sigma \implies A^{(k)} \in \Sigma$. By definition, Σ only contains $\mathbb{Q}$-linearly independent subsets of V, so $A^{(k)}$ must be $\mathbb{Q}$-linearly independent. This contradicts the law of excluded middle, so our assumption that $A := \bigcup_{i \in I} A_i$ is $\mathbb{Q}$-linearly dependent must be false. That is, we have shown that A is $\mathbb{Q}$-linearly independent. Since we already showed that A is a subset of V, this completes the proof that $A \in \Sigma$ by the definition of Σ. □

Since $A \in \Sigma$, and $A_i \subseteq A$ for all $i \in I$, we know $A := \bigcup_{i \in I} A_i$ is an upper bound to any chain $C = \{A_i | i \in I\} \subseteq \Sigma$.

Note that $(\Sigma, \subseteq)$ is nonempty for any $\mathbb{Q}$-vector space V since $\{v\}$ is $\mathbb{Q}$-linearly independent for all $v \in V$ and no $\mathbb{Q}$-vector space can be empty since it must contain the zero vector. Since $(\Sigma, \subseteq)$ is a nonempty poset in which every chain has an upper bound, Zorn's Lemma (Lemma 1.29 from the notes) guarantees the existence of a maximal element $M \in \Sigma$. That is, $\exists M \in \Sigma$ such that

$$M \subseteq A \implies A = M$$

for all $A \in \Sigma$.
Claim: M is a basis for V.

Proof. Assume to the contrary that M is not a basis for V. That is, there exists some $v \in V$ such that for all $k \in \mathbb{N}$, $m_1, ..., m_k \in M$, $\lambda_1, \cdots, \lambda_k \in \mathbb{Q} \setminus \{0\}$,

$$v \neq \sum_{i=1}^{k} \lambda_i m_i$$

If there exists $k \in \mathbb{N}$, $m_1, ..., m_k \in M$, $\lambda, \lambda_1, ..., \lambda_k \in \mathbb{Q} \setminus \{0\}$ such that

$$\sum_{i=1}^{k} \lambda_i m_i = -\lambda v$$

then

$$\sum_{i=1}^{k} (-\lambda)^{-1} \lambda_i m_i = (-\lambda)^{-1} \sum_{i=1}^{k} \lambda_i m_i = (-\lambda)^{-1}(-\lambda v) = ((-\lambda)^{-1} \cdot -\lambda) \cdot v = 1 \cdot v = v$$

which contradicts the law of excluded middle by assumption. Thus, we know that for all $k \in \mathbb{N}$, $m_1, ..., m_k \in M$, $\lambda, \lambda_1, ..., \lambda_k \in \mathbb{Q} \setminus \{0\}$,

$$\sum_{i=1}^{k} \lambda_i m_i \neq -\lambda v$$

Adding $-\lambda v$ to both sides yields

$$\lambda v + \sum_{i=1}^{k} \lambda_i m_i \neq -\lambda v + \lambda v \underbrace{=}_{2.c} 0$$

That is, there is no way to *non-trivially* express 0 as a rational linear combination of finitely many vectors in $M \cup \{v\}$. That is, $M \cup \{v\}$ is $\mathbb{Q}$-linearly independent. Since $v \in V$ and $M \subseteq V$ by definition, we know

$$M \cup \{v\} \subseteq V$$

So $M \cup \{v\}$ is a $\mathbb{Q}$-linearly independent subset of V, so $M \cup \{v\} \in \Sigma$. Clearly, for all $m \in M$, $m \in M \cup \{v\}$, so

$$M \subseteq M \cup \{v\}$$

However, $v \notin M$ since

$$\sum_{i=1}^{k} \lambda_i m_i \neq v$$

for all $k \in \mathbb{N}$, $m_1, ..., m_k \in M$, and $\lambda_1, ..., \lambda_k \in \mathbb{Q} \setminus \{0\}$ ($v \in M$ implies $k = 1$, $m_1 = v$, $\lambda_1 = 1$ works). Thus,

$$M \neq M \cup \{v\}$$

In summary, $M \cup \{v\} \in \Sigma$ and $M \subseteq M \cup \{v\}$ but $M \neq M \cup \{v\}$, which directly contradicts the maximality of M. By contradiction, this completes the proof that the maximal element $M \in \Sigma$ is a basis for V. □

Thus, for every $\mathbb{Q}$-vector space V, there exists a maximal element M in the set of all $\mathbb{Q}$-linearly independent subsets of V such that M is a basis for V. This completes the proof that every $\mathbb{Q}$-vector space has a basis.

Problem 4

(5 points). Consider the set of functions $V := Fun(\mathbb{N}, \mathbb{R})$ with $+ : V \times V \to V$ given by defining $f + g \in V$ to be the function

$$\forall n \in \mathbb{N}, \quad (f + g)(n) = f(n) + g(n)$$

and scalar multiplication $Q \times V \to V$ given by defining for each $\lambda \in \mathbb{Q}$ and $f \in V$, the function

$$\forall n \in \mathbb{N}, \quad (\lambda \cdot f)(n) = \lambda(f(n)).$$

Show that V is a $\mathbb{Q}$-vector space. Show that there is no countable basis for this vector space.

Hint: Is $\mathbb{R}$ a $\mathbb{Q}$-vector space? If so, does it have a countable basis? Note that every element in a $\mathbb{Q}$-vector space is a $\mathbb{Q}$-linear combination of *finitely* many basis elements.

Solution

First, we will show V is a $\mathbb{Q}$-vector space. It suffices to show V satisfies $(+_1), ..., (+_4)$ and $(\cdot_1), ..., (\cdot_4)$.

$(+_1)$ Consider the function $f_0 : \mathbb{N} \to \mathbb{R}$ defined by

$$f(n) = 0 \quad \forall n \in \mathbb{N}$$

Note that $f_0 \in Fun(\mathbb{N}, \mathbb{R}) =: V$. Then for all $f \in V$, $f + f_0, f_0 + f \in V$ satisfy

$$(f + f_0)(n) = f(n) + f_0(n) = f(n) + 0 = f(n) = 0 + f(n) = f_0(n) + f(n) = (f_0 + f)(n) \quad \forall n \in \mathbb{N}$$

since $0 + x = x = x + 0$ for all $x \in \mathbb{R}$ by definition of $0 \in \mathbb{R}$. That is,

$$f + f_0 = f = f_0 + f$$

This completes the proof that there is a zero vector $f_0 \in V$.

$(+_2)$ For all $f, g, h \in V$, $f + (g + h) \in V$ and $(f + g) + h \in V$ satisfy

$$(f+(g+h))(n) = f(n)+(g+h)(n) = f(n)+(g(n)+h(n)) = (f(n)+g(n))+h(n) = (f+g)(n)+h(n) = ((f+g)+h)(n)$$

for all $n \in \mathbb{N}$. Note that every equality follows from the definition of addition from the problem statement, except for the third equality, which follows from the associativity of addition in $\mathbb{R}$. That is, for all $f, g, h \in V$

$$f + (g + h) = (f + g) + h$$

which completes the proof that $+$, as defined in the problem statement, is associative.

$(+_3)$ Note that for all $f \in V$, since $f : \mathbb{N} \to \mathbb{R}$, for all $n \in \mathbb{N}$, $f(n) \in \mathbb{R}$ has an additive inverse. In lecture, we showed this additive inverse is $-f(n)$ (Lemma 2.9). That is, we know

$$f(n) + -f(n) = 0$$

for all $n \in \mathbb{N}$ and all $f \in V$. Thus, for all $f \in V$, define the function $f^- \in V$ such that

$$f^-(n) = -f(n) \quad \forall n \in \mathbb{N}$$

Then $f + f^- \in V$ satisfies

$$(f + f^-)(n) = f(n) + f^-(n) = f(n) + -f(n) = 0 = f_0(n)$$

for all $n \in \mathbb{N}$. Note that the first equality follows by the definition of $+$ from the problem statement, while the second follows from the definition of f^-, and the third follows from Lemma 2.9 in the notes. Thus, for all $f \in V$, we can find a $f^- \in V$ such that

$$f + f^- = f_0$$

This completes the proof that additive inverses exist.

($+_4$) In lecture, we took as an axiom that addition in $\mathbb{R}$ is commutative. That is,

$$x + y = y + x \quad \forall x, y \in \mathbb{R}$$

Thus, for all $f, g \in V$, $f + g$ and $g + f$ satisfy

$$(f + g)(n) = f(n) + g(n) = g(n) + f(n) = (g + f)(n)$$

for all $n \in \mathbb{N}$. Here, the first and third equalities follow by the definition of $+$ from the problem, statement, while the second equality follows from the commutativity of addition in $\mathbb{R}$. Thus, for all $f, g \in V$,

$$f + g = g + f$$

which completes the proof that $+$, as defined in the problem statement, is commutative.

($\cdot_1$) For all $\lambda \in \mathbb{Q}$, $f, g \in V$, $\lambda(f + g)$ satisfies

$$\lambda(f + g)(n) = \lambda(f(n) + g(n)) = \lambda f(n) + \lambda g(n) = (\lambda f)(n) + (\lambda g)(n)$$

for all $n \in \mathbb{N}$. The first equality follows by the definition of $+$ from the problem statement, while the second follows from the distributivity of multiplication over addition in $\mathbb{R}$ (which we took as an axiom in lecture), and the third equality follows from the definition of scalar multiplication from the problem statement. Thus, for all $\lambda \in \mathbb{Q}$, $f, g \in V$,

$$\lambda(f + g) = \lambda f + \lambda g$$

which completes the proof ($\cdot_1$).

($\cdot_2$) For all $\lambda, \mu \in \mathbb{Q}$, $f \in V$, $(\lambda + \mu)f$ satisfies

$$(\lambda + \mu)f(n) = \lambda f(n) + \mu f(n) = (\lambda f)(n) + (\mu f)(n)$$

for all $n \in \mathbb{N}$. The first equality follows by the distributivity of multiplication over addition in $\mathbb{R}$, while the second equality follows from the definition of scalar multiplication from the problem statement. Thus, for all $\lambda, \mu \in \mathbb{Q}$, $f \in V$, we have

$$(\lambda + \mu)f = \lambda f + \mu f$$

which completes the proof of ($\cdot_2$). Note that this combines with the previous proof to complete the proof of the distributivity of multiplication over addition in V.

($\cdot_3$) For all $\lambda, \mu \in \mathbb{Q}$, $f \in V$, we have

$$\lambda((\mu f)(n)) = \lambda(\mu f(n)) = (\lambda\mu)f(n)$$

for all $n \in \mathbb{N}$. The first equality follows by the definition of scalar multiplication from the problem statement, while the second follows from the commutativity of multiplication in $\mathbb{R}$. Thus, for all $\lambda, \mu \in \mathbb{Q}$, $f \in V$,

$$\lambda(\mu f) = (\lambda\mu)f$$

which completes the proof of ($\cdot_3$).

($\cdot_4$) For all $v \in V$, note that

$$(1 \cdot f)(n) = 1 \cdot f(n) = f(n)$$

for all $n \in \mathbb{N}$. The first equality follows from the definition of scalar multiplication from the problem statement, while the second follows from the axiom that 1, as the unit of the group $(\mathbb{R}, \cdot)$, satisfies $1 \cdot x = x$ for all $x \in \mathbb{R}$. Thus, for all $f \in V$, we have

$$1 \cdot f = f$$

which completes the proof that 1 acts trivially on all $f \in V$.

Since V satisfies $(+_1), (+_2), (+_3), (+_4)$ and $(\cdot_1), (\cdot_2), (\cdot_3), (\cdot_4)$, V is, by definition, a $\mathbb{Q}$-vector space under $+$ and $\cdot$ as defined in the problem statement.

It remains to show that V cannot have a countable basis.
Claim: If V is a $\mathbb{Q}$-vector space with countable basis B, then V itself must be countable.

Proof. Since B is countable, we can enumerate its elements. That is,

$$B = \{b_1, b_2, b_3, \ldots\}$$

where $b_i \in V$ for all $i \in \mathbb{N}$. Consider the sequence $B_1, B_2, B_3, ...$ defined by

$$B_i = \{b_1, b_2, \ldots, b_i\}$$

for all $i \in \mathbb{N}$. Note that, for any fixed $i \in \mathbb{N}$, B_i is a finite set of i vectors from V. Now, for all $i \in \mathbb{N}$, define V_i to be the $\mathbb{Q}$-vector space with basis B_i. That is,

$$V_i = \left\{ v \middle| v = \sum_{j=1}^{i} \lambda_j b_j, \lambda_j \in \mathbb{Q}, b_j \in B_i \right\}$$

Define $g_i : V_i \to \mathbb{Q}^i$ such that, for all $v \in V_i$,

$$g(v) = (\lambda_1, ..., \lambda_i) \text{ s.t. } v = \sum_{i=1}^{i} \lambda_j b_j$$

By definition of V_i, we know we can find such $(\lambda_1, ..., \lambda_i) \in \mathbb{Q}^i$ for all $v \in V_i$. Since B_i is fixed for fixed $i \in \mathbb{N}$, for any $v_1, v_2 \in V_i$,

$$g_i(v_1) = g_i(v_2) = (\lambda_1, ..., \lambda_i) \implies v_1 = \sum_{j=1}^{i} \lambda_j b_j = v_2$$

That is, for any fixed $i \in \mathbb{N}$, g_i is an injection from V_i to $\mathbb{Q}^i$, so

$$|V_i| \leq |\mathbb{Q}^i|$$

From Lemma 1.18 in the notes, we know $|\mathbb{Q}^i| = |\mathbb{N}|$ ($\mathbb{Q}^i$ is also countable as the finite product of the countable set $\mathbb{Q}$ by Proposition 1.7), so

$$|V_i| \leq |\mathbb{N}|$$

for any fixed $i \in \mathbb{N}$. By definition of countability, this means V_i is countable for any fixed $i \in \mathbb{N}$.

Claim: $V = \bigcup_{i \in \mathbb{N}} V_i$.

Proof. By the axiom of extensionality, it suffices to show

$$\forall v((v \in V) \iff (v \in \bigcup_{i \in \mathbb{N}} V_i))$$

First, suppose $v \in \bigcup_{i \in \mathbb{N}} V_i$. Then there exists some $i \in \mathbb{N}$ for which $v \in V_i$. By definition of V_i, there exist some $\lambda_1, ..., \lambda_i \in \mathbb{Q}$ such that

$$\sum_{j=1}^{i} \lambda_i b_i = v$$

That is, v can be expressed as a rational linear combination of the finitely many basis vectors $b_1, ..., b_i \in V_i$. Since

$$B_i := \{b_1, b_2, \ldots, b_i\} \subseteq \{b_1, b_2, \ldots\} =: B$$

v can also be expressed as the exact same rational linear combination of the finitely many vectors $b_1, ..., b_i \in B$, the basis of V. Thus, $v \in span(B)$, so $v \in V$ by definition of the basis. This completes the proof that

$$\forall v((v \in \bigcup_{i \in \mathbb{N}} V_i) \implies (v \in V))$$

Now, suppose $v \in V$. Then v must be a rational linear combination of finitely many basis vectors $b_1, ..., b_i \in B$. That is,

$$v = \sum_{j=1}^{i} \lambda_j b_j$$

for some $\lambda_1, ..., \lambda_i \in \mathbb{Q}$. By definition, $b_1, ..., b_i \in B_i := \{b_1, ..., b_i\}$. Thus, v can also be expressed as the same rational linear combination of finitely many basis vectors $b_1, ..., b_i \in B_i$, the basis of V_i. Thus, $v \in span(B_i)$, so $v \in V_i$. Since $V_i \subseteq \bigcup_{i \in \mathbb{N}} V_i$ by definition of the set union, this implies $v \in \bigcup_{i \in \mathbb{N}} V_i$. This completes the proof that

$$\forall v((v \in V) \implies (v \in \bigcup_{i \in \mathbb{N}} V_i))$$

and combines with the previous proof that

$$\forall v((v \in \bigcup_{i \in \mathbb{N}} V_i) \implies (v \in V))$$

to complete the proof that $V = \bigcup_{i \in \mathbb{N}} V_i$. □

Recall that V_i is countable for any fixed $i \in \mathbb{N}$. Therefore, Proposition 1.5 guarantees that

$$\bigcup_{i \in \mathbb{N}} V_i$$

must also be countable, as $\mathbb{N}$ is countable, so $\bigcup_{i \in \mathbb{N}} V_i$ is the countable union of countable sets. Since $V = \bigcup_{i \in \mathbb{N}} V_i$, this completes the proof that, if V is a $\mathbb{Q}$-vector space with countable basis B, then V itself must be countable. □

Now, we can finally prove $V := Fun(\mathbb{N}, \mathbb{R})$ has no countable basis. Assume to the contrary that V has a countable basis B. Then V must be countable by the previous result. Consider $h : \mathbb{R} \to Fun(\mathbb{N}, \mathbb{R})$ defined by

$$h(r) = f_r \in Fun(\mathbb{N}, \mathbb{R})$$

such that $f_r(n) = r$ for all $n \in \mathbb{N}$. Then, for any $r_1, r_2 \in \mathbb{R}$,

$$h(r_1) = h(r_2) \implies f_{r_1} = f_{r_2} \implies r_1 =: f_{r_1}(n) = f_{r_2}(n) := r_2$$

for all $n \in \mathbb{N}$. Thus,

$$(h(r_1) = h(r_2)) \implies r_1 = r_2$$

for all $r_1, r_2 \in \mathbb{R}$, so h is an injection from $\mathbb{R}$ to $Fun(\mathbb{N}, \mathbb{R})$. This implies

$$|\mathbb{R}| \leq |Fun(\mathbb{N}, \mathbb{R})| =: |V|$$

But we know from Theorem 1.21 in the notes that $\mathbb{R}$ is uncountable, so

$$|\mathbb{R}| > |\mathbb{N}| \implies |V| > |\mathbb{N}|$$

By the definition of uncountability, this means V must be uncountable, which contradicts the law of excluded middle since we already showed V must be countable due to its countable basis. By contradiction, the conclusion that V *cannot* have a countable basis follows.

Problem 5

(Bonus, 5 points). Fix $k \in \mathbb{N}_{\geq 2} := \{n \in \mathbb{N} | n \geq 2\}$. Prove that for every such k, the set

$$\Sigma_k := \{\phi : \mathbb{N} \to \mathbb{N} | \phi^{\circ k} = id_{\mathbb{N}}\}$$

is uncountable.

Hint: Start by considering the uncountable set constructed in problem 1.

Solution

From problem 1, we know

$$S := \{f \in \{0,1\}^{\mathbb{N}} | f^{-1}(0) \text{ and } f^{-1}(1) \text{ are both countably infinite}\}$$

is uncountable. Since both $f^{-1}(0)$ and $f^{-1}(1)$ are countably infinite for all $f \in S$, we can write

$$f^{-1}(0) = A_f = \{a_1, a_2, a_3, \ldots\}$$

and

$$f^{-1}(1) = B_f = \{b_1, b_2, b_3, \ldots\}$$

so that

$$A_f \cup B_f = f^{-1}(0) \cup f^{-1}(1) = f^{-1}(\{0,1\}) = \mathbb{N}$$

Now, fix $k \in \mathbb{N}_{\geq 2}$, and define $\phi_k : S \to \Sigma_k$ defined by

$$\phi_k(f) = g_f$$

for all $f \in S$, where

$$g_f(n) = \begin{cases} a_{i+1} \text{ if } n = a_i \text{ for some } i \in \mathbb{N} \text{ s.t. } i \not\equiv 0 \text{ (mod k)} \\ a_{i+1-k} \text{ if } n = a_i \text{ for some } i \in \mathbb{N} \text{ s.t. } i \equiv 0 \text{ (mod k)} \\ n \text{ if } n = b_i \text{ for some } i \in \mathbb{N} \end{cases}$$

for all $n \in \mathbb{N}$. Note that, for all $i \in \mathbb{N}$, $i+1 \in \mathbb{N}$. Also, for all $i \in \mathbb{N}$, $i \equiv 0$ (mod k) implies $i \geq k$, so $i+1-k \geq k+1-k = 1 \in \mathbb{N}$. Thus, for all $i \in \mathbb{N}$, g_f maps a_i to some distinct a_j, where $j \in \mathbb{N}$. Since $a_i \in \mathbb{N}$ for all $i \in \mathbb{N}$ by definition of $f \in S$ (and $n \in \mathbb{N}$ trivially for all $n \in \mathbb{N}$), this implies $g_f(n) \in \mathbb{N}$ for all $n \in \mathbb{N}$ and all $f \in S$. That, is $g_f : \mathbb{N} \to \mathbb{N}$ is well defined for all $f \in S$. Thus, to show $g_f \in \Sigma_k$ for all $f \in S$, it suffices to show $g_f^{\circ k} = id_{\mathbb{N}}$ for all $f \in S$. For all $f \in S$ and all $n \in B_f$ (i.e. $n = b_i$ for some $i \in \mathbb{N}$),

$$g_f^{\circ k}(n) = (\underbrace{g_f \circ \cdots \circ g_f}_{k \ g_f\text{'s}})(n) = g_f(g_f(\cdots g_f(g_f(n)))) = g_f(g_f(\cdots g_f(n))) = \cdots = g_f(g_f(n)) = g_f(n) = n$$

For all $n = a_i$, $i \in \mathbb{N}$, note that $g_f(n) \in \{a_{i+1}, a_{i+1-k}\}$. For all $i \in \mathbb{N}$, $i+1-k \equiv i+1 \not\equiv i$ (mod k), so $g_f(n) = a_j$ for some $j \in \mathbb{N}$ such that $j \neq i$ and $j \equiv i+1$ (mod k). By induction, when $g_f^{\circ k}(a_i)$ is evaluated, the index l of the natural number $a_l \in A_f$ being plugged into g_f satisfies $l \equiv j$ (mod k) exactly once for all $j \in \{0, 1, ..., k-1\}$. By definition of g_f, this means the initial index i is incremented by one $k-1$ times and incremented by $1-k$ once. If we let $g^{\circ k}(a_i) = a_{i_f}$, this implies

$$i_f = i + (1 \cdot (k-1)) + ((1-k) \cdot 1) = i + k - 1 + 1 - k = i$$

That is,

$$\begin{aligned} g^{\circ k}(n) = & \quad g_f(g_f(\cdots g_f(g_f(a_i)))) = g_f(g_f(\cdots g_f(a_{j \in \{i+1, i+1-k\}}))) \\ = \cdots = & \quad g_f(g_f(a_{j \in \{i+k-2, i-2\}})) = g_f(a_{j \in \{i+k-1, i-1\}}) = a_i = n \end{aligned}$$

for all $f \in S$ and all $n = a_i \in A_f := f^{-1}(0)$. Thus, for all $f \in S$,

$$g_f^{\circ k}(n) = n$$

for all $n \in \mathbb{N}$. By definition of the identity function, this implies

$$g_f^{\circ k} = id_{\mathbb{N}}$$

for all $f \in S$. Since our modulo arithmetic holds for all $k \in \mathbb{N}_{\geq 2}$, this implies $\phi_k : S \to \Sigma_k$ is well-defined for all such k.

We now focus on showing that ϕ_k is an injection from S to Σ_k for all $k \in \mathbb{N}_{\geq 2}$. If $g_{f_1} = g_{f_2}$ for some $f_1, f_2 \in S$, then for all $a_i \in A_{f_1} := \{a_1, a_2, \ldots\} := f_1^{-1}(0)$, we have

$$g_{f_1}(a_i) = a_j \text{ s.t. } j \neq i$$

so $g_{f_1}(a_i) \neq a_i$. If $a_i \in B_{f_2} := f_2^{-1}(1)$, then

$$g_{f_2}(a_i) = a_i$$

by definition of g. But, $g_{f_1} = g_{f_2}$, so

$$a_i \neq a_j = g_{f_1}(a_i) = g_{f_2}(a_i) = a_i$$

which contradicts the law of excluded middle since $a_i = a_i$ is trivially true. Thus, for all $a_i \in A_{f_1}$, $a_i \in \mathbb{N} \setminus B_{f_2} = A_{f_2}$. Similarly, for all $b \in B_{f_1}$, we have

$$g_{f_1}(b) = b$$

If $b = a_i \in A_{f_2}$, then

$$g_{f_2}(b) = a_j \text{ for some } j \in \mathbb{N} \text{ s.t. } a_j \neq a_i$$

But $g_{f_1} = g_{f_2}$, so

$$b = a_i \neq a_j = g_{f_2}(b) = g_{f_1}(b) = b$$

which once again contradicts the law of excluded middle. Thus, for all $b \in B_{f_1}$, $b \in \mathbb{N} \setminus A_{f_2} = B_{f_2}$.
The inclusions hold in the opposite direction too. For all $a_i \in A_{f_2}$,

$$g_{f_2}(a_i) = a_j \text{ for some } j \in \mathbb{N} \text{ s.t. } a_i \neq a_j$$

If $a_i \in B_{f_1}$, then

$$g_{f_1}(a_i) = a_i$$

but $g_{f_1} = g_{f_2}$, so

$$a_i \neq a_j = g_{f_2(a_i)} = g_{f_1}(a_i) = a_i$$

which contradicts the law of excluded middle. Thus, for all $a_i \in A_{f_2}$, $a_i \in A_{f_1}$. Similarly, for all $b \in B_{f_2}$, we have

$$g_{f_2}(b) = b$$

If $b = a_i \in A_{f_1}$, then

$$g_{f_1}(b) = g_{f_1}(a_i) = a_j \text{ for some } j \in \mathbb{N} \text{ s.t. } a_j \neq a_i$$

but $g_{f_1} = g_{f_2}$, which implies

$$b = a_i \neq a_j = g_{f_1}(b) = g_{f_2}(b) = b$$

which once again yields a contradiction with the law of excluded middle. Thus, for all $b \in B_{f_2}$, $b \in \mathbb{N} \setminus A_{f_1} = B_{f_1}$. Combining these results yields

$$\forall a((a \in A_{f_1}) \iff (a \in A_{f_2})) \quad \text{and} \quad \forall b((b \in B_{f_1}) \iff (b \in B_{f_2}))$$

By the axiom of extensionality, this implies $A_{f_1} = A_{f_2}$ and $B_{f_1} = B_{f_2}$. Thus, for all $n \in \mathbb{N}$,

$$f_1(n) = 0 \iff f_2(n) = 0 \quad \text{and } f_1(n) = 1 \iff f_2(n) = 1$$

That is,

$$\phi(f_1) := g_{f_1} = g_{f_2} =: \phi(f_2) \implies f_1 = f_2$$

so ϕ_k is an injection from S to Σ_k, and

$$|S| \leq |\Sigma_k|$$

From **Problem 1**, we know that S is uncountable, so

$$|\mathbb{N}| < |S|$$

Combining our inequalities yields

$$|\mathbb{N}| < |S| \leq |\Sigma_k| \implies |\mathbb{N}| < |\Sigma_k|$$

for all $k \in \mathbb{N}_{\geq 2}$. By the definition of countability, this completes the proof that Σ_k is uncountable for all $k \in \mathbb{N}_{\geq 2}$.

MATH 458: Numerical Methods

All assignmetns in this section were written by Aykut Arslan, Lecturer, USC. Solutions to assignments 1 and 2 are provided.

Assignment 1

Problem 1

The two links below have descriptions of four different situations in which errors arose in numerical computing with serious consequences. Choose *one* of these situations (or another situation of your choice and provide a reference) and write a short paragraph (a few sentences will suffice) about what the error was and what happened.

- `https://en.wikipedia.org/wiki/Pentium_FDIV_bug`
- `https://www.iro.umontreal.ca/~mignotte/IFT2425/Disasters.html`

Solution

The Patriot Missile failure highlights the danger of allowing small rounding errors to accumulate. The Patriot Missile battery maintains an internal clock that stores time since boot to the nearest second. In order to make accurate physical calculations, the battery frequently multiplies this clock value by $\frac{1}{10}$ to obtain the time since boot in tenths of a second. Unfortunately, the decimal number $\frac{1}{10} = 0.1$ cannot be expressed exactly as a finite binary number, as $0.1_{10} = 0.0\overline{0011}_2$, which is a non-terminating sequence of bits. The Patriot Missile system utilizes only 24 bits for the register storing local time, which introduces about $9.5 \cdot 10^{-8}$ rounding error per tenth of a second that the battery has been running. While insignificant when the battery has low up-time, this error accumulates dangerously when the system runs for a long period of time. In 1991, after running for around 100 hours straight, a Patriot Missile battery in Saudi Arabia missed an Iraqi Scud missile which killed 28 soldiers in an American Army barracks. The 100 hour up-time led to an accumulated rounding error of around $\frac{1}{3}$ of a second, enough time for the incoming missile to change position by more than half a kilometer, rendering the missile defense system ineffective in this situation. The consequences of this failure demonstrate the importance of understanding how small errors accumulate with extreme input values.

Problem 2

Consider a polynomial

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

(a) Given a value of x, how many multiplications and how many additions are needed to calculate f(x) using the naive formula above? *Hint: For example, x^4 is calculated as $x \cdot x \cdot x \cdot x$, so it requires 3 multiplications.*

(b) Are the total number of operations (additions and multiplications combined) in (a), $\mathcal{O}(n)$, $\mathcal{O}(n^2)$, or $\mathcal{O}(n^3)$?

(c) What if we first calculate the powers of x are store them:

$$\begin{aligned} x_2 &= x \cdot x \\ x_3 &= x \cdot x_2 \\ x_4 &= x \cdot x_3 \\ &\vdots \end{aligned}$$

and then evaluate the function as

$$f(x) = a_n x_n + \cdots + a_2 x_2 + a_1 x + a_0?$$

Now, how many multiplications and additions are performed? Is the total number of operations $\mathcal{O}(n)$, $\mathcal{O}(n^2)$, or $\mathcal{O}(n^3)$?

(d) We can also write the polynomial in nested form as

$$f(x) = (\cdots(((a_n x + a_{n-1})x + a_{n-2})x + a_{n-3})x + \cdots + a_1)x + a_0$$

How many multiplications and how many additions are used when this form is used? Is it $\mathcal{O}(n)$, $\mathcal{O}(n^2)$, or $\mathcal{O}(n^3)$?

Solution

(a) Let $N :=$ the number of addition operations needed to compute the naive formula of $f(x)$, and let $M :=$ the number of multiplication operations needed to compute the naive formula of $f(x)$.
Note that the naive formula is a sum of $n+1$ terms (one for each power of x in $\{0, 1, ..., n\}$).
Claim: A sum of k terms takes $k-1$ addition operations to compute, for all $k \in \{1, ..., n+1\}$.
Proof. We induct on k.
Base Case: When $k = 1$, we have a sum of a single term, which takes $0 = k - 1$ addition operations to compute.
Inductive Hypothesis: Assume a sum of k terms takes $k-1$ addition operations to compute for all $1 \leq k \leq i < n+1$.
Inductive Step: Consider $k = i+1$. A sum of $i+1$ terms is just a sum of i terms plus one more term. From the inductive hypothesis, we know a sum of i terms takes $i-1$ addition operations to compute. We need one more addition operation to compute our sum of $i+1$ terms, so we need $i-1+1 = i = k-1$ addition operations total.
The conclusion that a sum of k terms takes $k-1$ addition operations to compute follows by induction for all $k \in \{1, ..., n\}$.
This result directly implies that the sum of $n+1$ terms in the naive formula for the polynomial will take exactly n addition operations to compute.

Thus, we can just calculate the number of multiplication operations for each of the $n+1$ terms, then add them up. *Claim:* x^i takes $i-1$ multiplication operations to compute, for all $i \in \{1, ..., n\}$.
Proof. We induct on i.
Base Case: When $i = 1$, $x^i = x$, which takes $0 = i - 1$ multiplication operations to compute, so the claim holds for the base case.
Inductive Hypothesis: Assume x^i takes $i-1$ multiplication operations to compute, for all $1 \leq i \leq k < n$.
Inductive Step: Consider $i = k+1$. Then $x^i = x^{k+1} = x^k \cdot x$. From the inductive hypothesis, we know x^k takes $k-1$ multiplication operations to compute. We need one more multiplication operation to compute $x^{k+1} = x^k \cdot x$, so we have a total of $k - 1 + 1 = k$ multiplication operations.
The conclusion that x^i takes $i-1$ multiplication operations to compute follows by induction for all $i \in \{1, ..., n\}$.
We use the previous result to calculate the number of multiplication operations needed to compute each of the $n+1$ terms in the naive formula for $f(x)$.
Claim: The term with coefficient a_i takes a total of i multiplication operations to compute.
Proof. Note that, when $i = 0$, the term with a_i is just a_0, which takes $0 = i$ multiplication operations to compute. For all $i \in \{1, ..., n\}$, the term with a_i is of the form $a_i x^i$. From the previous proof, we know x^i takes $i-1$ multiplication operations to compute. We need one more multiplication operation to get $a_i \cdot x_i$, so it takes $i-1+1 = i$ multiplication operations to compute. Thus, for all $i \in \{0, 1, ..., n\}$, the term with a_i takes i multiplication operations to compute.
Summing the number of multiplication operations to compute the term with coefficient a_i for all $i \in \{0, 1, ..., n\}$, we find

$$M = \sum_{i=0}^{n} i = \sum_{i=1}^{n} i \qquad (1)$$

Claim:

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \qquad (2)$$

for all $n \in \mathbb{N}$.
Proof. We induct on n.
Base Case: When $n = 1$, we have

$$\sum_{i=1}^{1} i = 1 = \frac{1(1+1)}{2} = \frac{2}{2} \tag{3}$$

so (2) holds in the base case.
Inductive Hypothesis: Assume (2) holds for all $1 \leq n \leq k$, $n \in \mathbb{N}$.
Inductive Step: Consider $n = k + 1$. Note that

$$\sum_{i=1}^{n} i = \sum_{i=1}^{k+1} i = (k+1) + \sum_{i=1}^{k} i \tag{4}$$

From the inductive hypothesis, we know that $\sum_{i=1}^{k} i = \frac{k(k+1)}{2}$. Plugging this into (4) yields

$$\sum_{i=1}^{n} i = (k+1) + \frac{k(k+1)}{2} = \frac{2(k+1) + k(k+1)}{2} = \frac{(k+2)(k+1)}{2} = \frac{n(n+1)}{2} \tag{5}$$

The conclusion that (2) holds for all $n \in \mathbb{N}$ follows from (5) by induction.
Plugging this result into (1) yields

$$M = \sum_{i=1}^{n} i = \frac{n(n+1)}{2} = \frac{n^2 + n}{2} \tag{6}$$

In summary, there are a total of $N = n$ addition operations and $M = \frac{n^2+n}{2}$ multiplication operations needed to compute the naive formula for $f(x)$.

(b) *Claim:* The total number of operations (additions and multiplications combined) in (a) is $\mathcal{O}(n^2)$.
Proof. Note that the total number of operations $T_a(n)$ needed to compute the naive formula for $f(x)$ is

$$T_a(n) = N + M = n + \frac{n(n+1)}{2} = \frac{2n + (n+1)n}{2} = \frac{(n+3)n}{2} \tag{7}$$

Thus, it suffices to find two constants $n_0 > 0, \in \mathbb{N}$ and $C > 0, \in \mathbb{R}$ such that

$$T_a(n) = \frac{(n+3)n}{2} \leq Cn^2 \tag{8}$$

for all $n \geq n_0$. We claim that $n_0 = 1$, $C = 10$ are sufficient. When $n = n_0$, we have

$$T_a(n) = T_a(1) = \frac{1(1+3)}{2} = \frac{4}{2} = 2 \leq 10(1)^2 = 10 \tag{9}$$

so we just need to show that $10n^2$ grows faster than $T_a(n)$ for all $n \geq 1$. Note that

$$\frac{d}{dn} T_a(n) = x + \frac{3}{2} \tag{10}$$

while

$$\frac{d}{dn} 10n^2 = 20n \tag{11}$$

Comparing (10) and (11), we see that $\frac{d}{dn} T_a(n) = \frac{5}{2} \leq 20 = \frac{d}{dn} 10n^2$ when $n = n_0$, and $\frac{d}{dn} T_a(n)$ grows 20 times slower than $\frac{d}{dn} 10n^2$ for all $n \geq n_0$. Thus, we can conclude

$$\frac{d}{dn} T_a(n) \leq \frac{d}{dn} 10n^2 \tag{12}$$

for all $n \geq n_0$. This combines with (9) to complete the proof that

$$T_a(n) \leq 10n^2 \qquad (13)$$

for all $n \geq 1 = n_0$, which completes the proof that the total number of operations (additions and multiplications combined) in (a) is

$$T_a(n) = \mathcal{O}(n^2) \qquad (14)$$

(c) Note that the optimized formula

$$f(x) = a_n x_n + \cdots + a_2 x_2 + a_1 x + a_0$$

is still a sum of $n+1$ terms, so it still requires $N = n$ addition operations to compute, by part (a). However, the number of multiplication operations required has changed. It still requires 0 and 1 multiplication operations to compute the terms with coefficient a_0 and a_1, respectively. However, for all $i \in \{2, ..., n\}$, $x_i := x^i = x \cdot x_{i-1}$, so each x^i only takes *one* multiplication operation to compute (assuming x_{i-1} has already been computed). Thus, for all $i \in \{2, ..., n\}$, the number of multiplication operations needed to compute the term $a_i x_i = a_i x^i$ is 2 (one for x_i and one for $a_i \cdot x_i$). Adding up the needed multiplication operations for all $n+1$ terms, we find

$$M' = 0 + 1 + 2(n-1) = 2n - 2 + 1 = 2n - 1 \qquad (15)$$

where $M' :=$ the number of multiplication operations needed to compute f(x) with the optimized formula from part (c). This follows because there are $n-1$ terms with coefficients a_i where $i \in \{2, ..., n\}$. Combining this with the unchanged $N = n$ number of addition operations needed, we find

$$T_b(n) = M' + N = 2n - 1 + n = 3n - 1 \qquad (16)$$

where $T_b(n) :=$ the number of total operations (addition and multiplication combined) needed to compute the optimized formula for $f(x)$.
Claim: The total number of operations performed to compute the optimized formula for $f(x)$ is $\mathcal{O}(n)$.
Proof: From (16) and the definition of Big-$\mathcal{O}$ notation, it suffices to find two constants $n_0 > 0, \in \mathbb{N}$ and $C > 0, \in \mathbb{R}$, such that

$$T_b(n) = 3n - 1 \leq Cn \qquad (17)$$

for all $n \geq n_0$. We will show that $n_0 = 1$, $C = 4$ are sufficient. Note that, when $n = n_0$,

$$T_b(n) = T_b(1) = 3(1) - 1 = 2 \leq 4(1) = 4 \qquad (18)$$

so it suffices to show $T_b(n)$ grows no faster than $4n$ for all $n \geq 1 = n_0$. Differentiating, we can easily see that

$$\frac{d}{dn} T_b(n) = 3 \leq 4 = \frac{d}{dn} 4n \qquad (19)$$

Thus, $T_b(n) \leq 4n$ when $n = n_0$ and $T_b(n)$ grows no faster than $4n$ for all $n \geq n_0$, allowing us to conclude that

$$T_b(n) \leq 4n \qquad (20)$$

for all $n \geq 1 = n_0$. The result from (20) completes the proof that the total number of operations (additions and multiplications combined) needed to compute the optimized polynomial formula is

$$T_b(n) = \mathcal{O}(n) \qquad (21)$$

(d) *Claim:* For any polynomial of degree $n \in \mathbb{N} \cup \{0\}$, it takes $N" = n$ addition and $M" = n$ multiplication operations to compute the nested formula for $f(x)$.
Proof. We induct on n. *Base Case:* When $n = 0$, the nested formula becomes

$$f(x) = (\cdots(((a_n x + a_{n-1})x + a_{n-2})x + a_{n-3})x + \cdots + a_1)x + a_0 = a_0 \qquad (22)$$

which requires $N" = n = 0$ addition operations and $M" = n = 0$ multiplication operations to compute.
Inductive Hypothesis: Assume that a polynomial of degree n takes $N" = n$ addition and $M" = n$ multiplication operations for all $0 \leq n \leq k$, $n, k \in \mathbb{N} \cup \{0\}$.
Inductive Step: Consider $n = k + 1$. We have the nested formula

$$f(x) = (\cdots(((a_{k+1}x + a_k)x + a_{k-1})x + a_{k-2})x + \cdots + a_1)x + a_0 \qquad (23)$$

For $n = k$, we have the nested formula

$$f(x) = (\cdots(((a_k x + a_{k-1})x + a_{k-2})x + a_{k-3})x + \cdots + a_1)x + a_0 \qquad (24)$$

Comparing (23) and (24), the only difference between the two formulae is that (23) has $(a_{k+1}x + a_k)$ where (24) just has a_k. Thus, the number of addition operations needed to compute (23) is just the number of addition operations needed to compute (24) plus the number of addition operations needed to compute $(a_{k+1}x + a_k)$. Similarly, the number of multiplication operations needed to compute (23) is just the number needed to compute (24) plus the number needed to compute $(a_{k+1}x + a_k)$. By the inductive hypothesis, it takes k addition operations and k multiplication operations to compute (24) and it only takes one addition operation and one multiplication operation to compute $(a_{k+1}x + a_k)$, so it takes

$$N" = k + 1 = n \qquad (25)$$

total addition operations and

$$M" = k + 1 = n \qquad (26)$$

total multiplication operations to compute (23).
The conclusion that it takes exactly $N" = n$ addition operations and $M" = n$ multiplication operations to compute the nested formula for $f(x)$ follows by induction from (25) and (26) for all $n \in \mathbb{N} \cup \{0\}$. Thus, the total number of operations (addition and multiplication combined) needed to compute the nested formula for $f(x)$ is

$$T_c(n) = M" + N" = n + n = 2n \qquad (27)$$

Claim: The total number of operations needed to compute the nested formula for $f(x)$ is $T_c(n) = \mathcal{O}(n)$.
Proof. It suffices to find two constants $n_0 > 0, \in \mathbb{N}$ and $C > 0, \in \mathbb{R}$, such that

$$T_c(n) = 2n \leq Cn \qquad (28)$$

for all $n \geq n_0$. Choosing $n_0 = 1$, $C = 2$, the statement from (28) becomes

$$T_c(n) = 2n \leq 2n \text{ for all } n \geq 1 \qquad (29)$$

which is a vacuously true statement (since $2n = 2n \implies 2n \leq 2n$ and $2n \geq 2n$ for all $n \in \mathbb{R}$). This completes the proof that the total number of operations (additions and multiplications combined) needed to compute the nested formula for $f(x)$ is

$$T_c(n) = \mathcal{O}(n) \qquad (30)$$

Problem 3

Consider the polynomial f that is given by

$$f(x) = x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1$$

(a) Plot the graph of this function for values of x from $x = 0.988$ to $x = 1.012$ using steps of size 0.0001. Show the computer code you used to do this and the plot you obtained when you submit your homework. Does this look like a polynomial?

(b) In fact $f(x) = (x-1)^7$ (check this!). Use this formula for f to plot the graph of f for the same values of x. This looks much better!

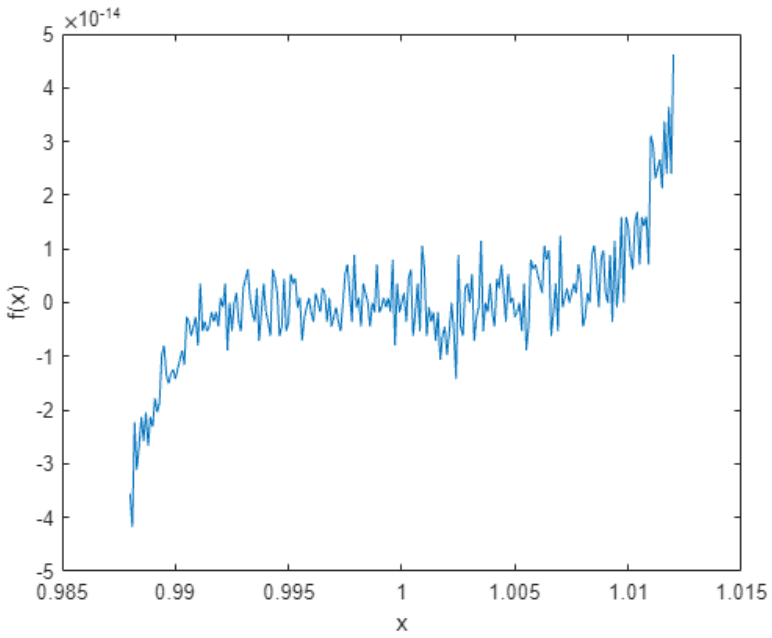
(c) Using the values calculated in (b) as the exact values (they aren't quite exact but they are pretty close), find the error and relative error when you calculate $f(x)$ using (a) and plot these errors as a function of x. Use a log scale on the y-axis for the graph of the relative error. (In Matlab you can use the command semilogy). Show your plots when you turn in your homework.

Solution

(a) We use the following MATLAB code:

```
x = 0.988:.0001:1.012;
y = x.^7 - 7.*x.^6 + 21.*x.^5 - 35.*x.^4 + 35.*x.^3 - 21.*x.^2 + 7.*x - 1;
plot(x, y)
xlabel("x");
ylabel("f(x)");
```

to produce the following plot



The plot is quite jagged, with many more than the 6 local extrema expected for a polynomial of degree 7. It thus does not look much like the degree 7 polynomial it is supposed to represent.

(b) We will use the Binomial Theorem to verify that $f(x) = (x-1)^7$. It states that

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i} \qquad (31)$$

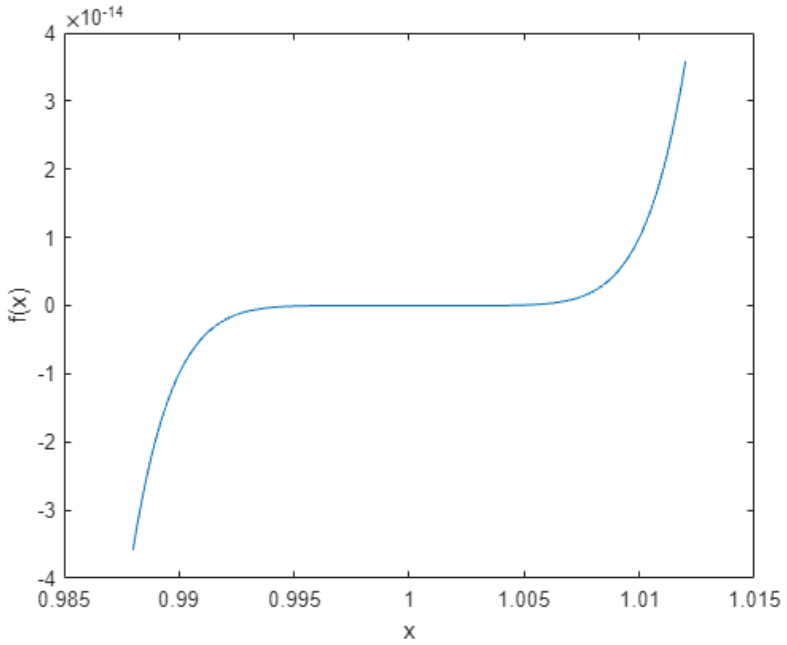
Applying (31) to $(x-1)^7$ yields

$$\begin{aligned}
(x-1)^7 &= \sum_{i=1}^{7} \binom{7}{i} x^i (-1)^{7-i} \\
&= -\binom{7}{0}x^0 + \binom{7}{1}x^1 - \binom{7}{2}x^2 + \binom{7}{3}x^3 - \binom{7}{4}x^4 + \binom{7}{5}x^5 - \binom{7}{6}x^6 + \binom{7}{7}x^7 \\
&= -1 + 7x - \frac{7!}{5!2!}x^2 + \frac{7!}{4!3!}x^3 - \frac{7!}{4!3!}x^4 + \frac{7!}{5!2!}x^5 - 7x^6 + x^7 \\
&= x^7 - 7x^6 + 21x^5 - 35x^4 + 35x^3 - 21x^2 + 7x - 1 \\
&= f(x) \qquad (32)
\end{aligned}$$

so $f(x) = (x-1)^7$ is indeed true. Having verified this equality, we use the following MATLAB code:

```
x2 = 0.988:.0001:1.012;
y2 = (x2- 1).^7;
plot(x2, y2)
xlabel("x");
ylabel("f(x)");
```

to produce the following plot for $f(x)$ using the new formula:

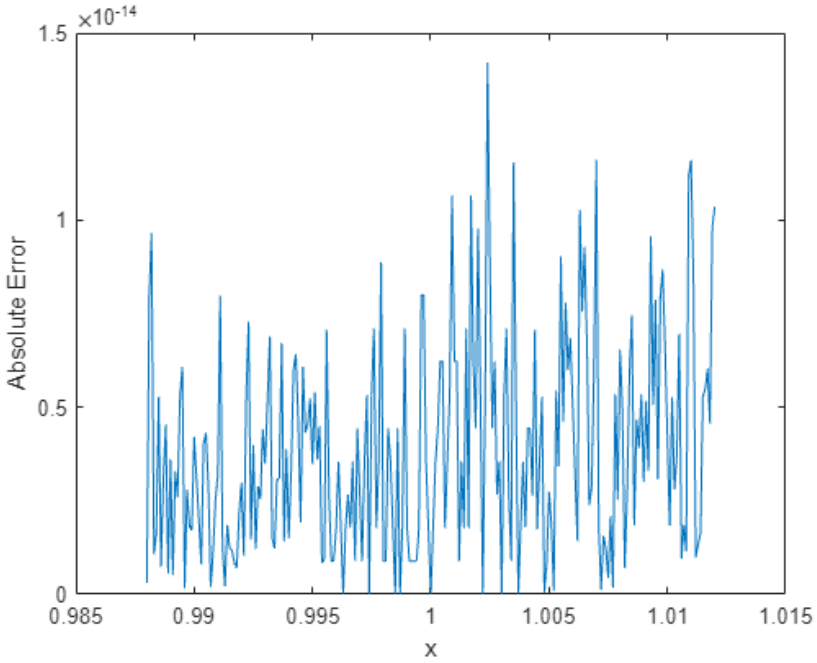


This plot is much smoother and looks much more like the expected plot for a polynomial in x.

(c) With $x2$, y, and $y2$ defined as in the MATLAB code from parts (a) and (b), we use the following MATLAB code:

```
ea = abs(y2 - y);
plot(x2, ea);
xlabel("x");
ylabel("Absolute Error");
```

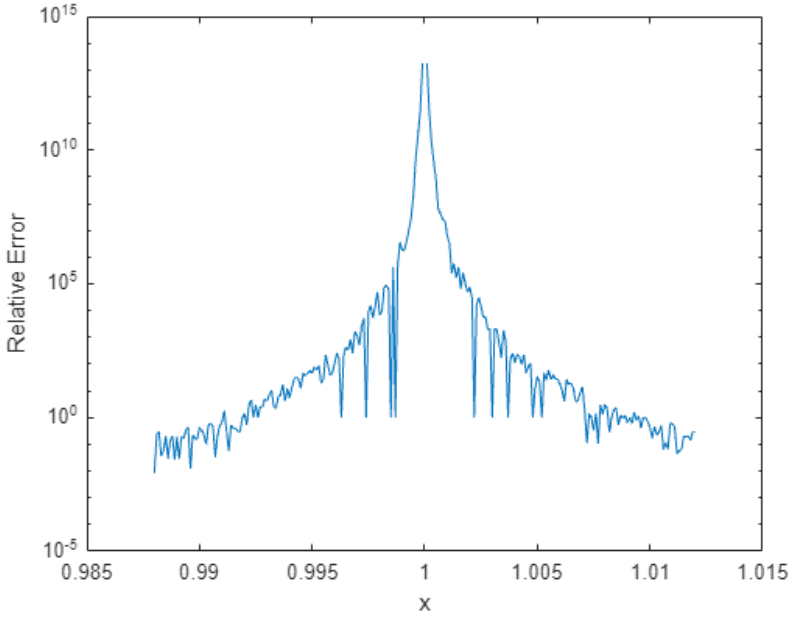
to produce the following plot for the absolute error of calculating $f(x)$ using the formula from (a):



For the relative error, we use the following MATLAB code:

```
er = abs((y2 - y)./y2);
semilogy(x2, er);
xlabel("x");
ylabel("Relative Error");
```

to produce the following plot:



Note that, unlike the plot for absolute error, this plot has a logarithmic scale on the y-axis.

Problem 4

Consider the problem of evaluating the integrals

$$y_n = \int_0^1 \frac{x^n}{x+10} dx$$

for $n = 1, 2, ...$ Analytically, we notice that

$$y_n + 10y_{n-1} = \int_0^1 \frac{x^n + 10x^{n-1}}{x+10} dx = \int_0^1 x^{n-1} dx = \frac{1}{n}.$$

Also,

$$y_0 = \int_0^1 \frac{1}{x+10} = ln(11) - ln(10)$$

This gives us the following algorithm for computing $y_0, y_1, ...$:

1. Compute $y_0 = ln(11) - ln(10)$.
2. For $n = 1, 2, ...$, evaluate $y_n = \frac{1}{n} - 10y_{n-1}$.

(a) Show that $y_n < 1$ for all n and y_n decreases monotonically to 0 as $n \to \infty$. (A picture would suffice for this!)

(b) Program the algorithm in a computer and find y_n for $n = 1, 2, ..., 30$. What happens? Why does this happen?

(c) Derive an algorithm for computing the values of these integrals based on evaluating the value of y_{n-1} from the value of y_n.

(d) Suppose you want to calculate the values of $y_0, y_1, ..., y_N$ for some number N with an absolute error less than ε for some $\varepsilon > 0$. Since the naive algorithm above produces large errors (even though it is exact in theory), you will use the algorithm in (c). But you need a starting value. Show that there exists $M \in \mathbb{N}$ such that if you take $y_M = 0$ and use the algorithm in (c), then, if calculations are performed with infinite precision, the absolute errors in the calculations of $y_0, y_1, ..., y_N$ will be less than ε.

(e) Explain why rounding errors in the computer do not produce excessive errors using this algorithm.

(f) Use your algorithm (and the computer) to find the value of y_{30} to an accuracy of 10^{-5}. Explain how you chose M in this case.

Solution

(a) To show that $y_n < 1$ for all n, we just need to show that the integrand $\frac{x^n}{x+10} < 1$ for all $x \in (0,1)$ (we can exclude the boundaries because a single point cannot contribute to the integral's final value) and all $n \in \mathbb{N}$ (we can exclude $n = 0$ since $y_0 = ln(11) - ln(10) \approx 0.095 < 1$). Note that $x^n < 1$ for all $x \in (0,1)$ and all $n \in \mathbb{N}$. Also, $x + 10 > 1$ for all $x \in (0,1)$ and all $n \in \mathbb{N}$. Thus, for all $n \in \mathbb{N}$ and all $x \in (0,1)$, we have

$$\frac{x^n}{x+10} < \frac{1}{x+10} < \frac{1}{1} = 1 \qquad (33)$$

The result from (33) directly implies that

$$\int_0^1 \frac{x^n}{x+10} dx < \int_0^1 1 dx = 1 \qquad (34)$$

for all $n \in \mathbb{N}$. Since $y_0 = ln(11) - ln(10) < 1$, we have

$$y_n < 1 \text{ for all } n \in \mathbb{N} \qquad (35)$$

To show that y_n decreases monotonically, we must show $y_n < y_{n-1}$ for all $n \in \mathbb{N}$. To do so, it suffices to show that the integrand $\frac{x^n}{x+10} < \frac{x^{n-1}}{x+10}$ for all $x \in (0,1)$ and $n \in \mathbb{N}$. This follows directly from the fact that

$$x^n = x^{n-1} \cdot x < x^{n-1} \cdot 1 = x^{n-1} \qquad (36)$$

for all $n \in \mathbb{N}$ and all $x \in (0,1)$. Thus, we have

$$y_n := \int_0^1 \frac{x^n}{x+10} dx < \int_0^1 \frac{x^{n-1}}{x+10} dx = y_{n-1} \qquad (37)$$

for all $n \in \mathbb{N}$.
To show that y_n converges to 0 as $n \to \infty$, it suffices to show that the integrand $\frac{x^n}{x+10}$ converges to 0 as $n \to \infty$ for all $x \in (0,1)$. Note that

$$\lim_{n\to\infty} \frac{x^n}{x+10} = \frac{1}{x+10} \lim_{n\to\infty} x^n = \frac{1}{x+10} \lim_{n\to\infty} e^{ln(x^n)} = \frac{1}{x+10} \lim_{n\to\infty} e^{nln(x)} \qquad (38)$$

For all $x \in (0,1)$, $ln(x) < 0$, so

$$\lim_{n\to\infty} nln(x) = -\infty \qquad (39)$$

Thus, we can rewrite (38) as

$$\lim_{n\to\infty} \frac{x^n}{x+10} = \frac{1}{x+10} \lim_{t\to-\infty} e^t = \frac{1}{x+10} \lim_{t\to\infty} \frac{1}{e^t} = \frac{1}{x+10} \cdot 0 = 0 \qquad (40)$$

for all $x \in (0,1)$, with the last equality following since $\frac{1}{x+10} > 0$ for all $x \in (0,1)$. From (40), we see that, as $n \to \infty$, we have

$$\lim_{n\to\infty} y_n := \lim_{n\to\infty} \int_0^1 \frac{x^n}{x+10} dx = \int_0^1 \lim_{n\to\infty} \frac{x^n}{x+10} dx = \int_0^1 0 dx = 0 \qquad (41)$$

with the second to last inequality following from the Monotone Convergence Theorem. This completes the proof that $y_n < 1$ for all $n \to \infty$ and y_n decreases monotonically to 0 as $n \to \infty$.

(b) We use the following MATLAB code:

```
y = zeros(30);
y0 = log(11) - log(10);
y(1) = 1 - 10*y0;
for i = 2:30
    y(i) = (1/i) - 10*y(i-1);
end
```

to produce the following values for $y_1, ..., y_{30}$:

$$\begin{aligned}(y_1, ..., y_{30}) = \quad & (0.046898, 0.03101798, 0.02315, 0.01846, 0.01535, 0.0131, 0.01148, 0.01019, 0.009167, \\ & 0.008327, 0.0076, 0.0069, 0.00745, -0.003074, 0.09741, -0.9116, 9.17, -91.69, 916.99, \\ & -9169.877, 91698.821, -916988.2, 9169881.699, -91698816.9, 916988169.5, -9169881695.32, \\ & 91698816953.3, -916988169532.8, 9169881695328.4, -91698816953283.7) \qquad (42)\end{aligned}$$

Comparing these results to the theoretical results from (a), we see that the computed values of y_n break the monotone decreasing nature of the sequence around y_{13}. Furthermore, the computed values do not satisfy the restriction that $|y_n| < 1$ for all $n \in \mathbb{N}$, and some of the y_n are computed to be negative, which is theoretically impossible. Thus, the computed values are clearly producing very high error using the naive algorithm provided above.
The computed values differ so much from the theoretical behavior of the sequence due to the accumulation of rounding error. Computing y_n for all $n \in \{1, ..., 30\}$ with the naive algorithm involves the computation of values like $\frac{1}{3}$, $\frac{1}{10}$, $ln(11)$, and $ln(10)$, none of which can be represented exactly by a terminating sequence of bits. Thus, rounding error is introduced for each of these terms when they are first computed. The final error is so significant because of how these rounding errors accumulate as y_n is computed for higher values of n. At iteration i, the (potential) rounding error of $\frac{1}{i}$ is added to the total rounding error for y_{i-1} multiplied by 10. That is, the rounding error introduced by a computation in iteration i (like $ln(11)$ for $i = 0$ or $\frac{1}{i}$ for $i > 0$) is multiplied by 10^{n-i} by the time the value of y_n is computed, for all $0 \leq i \leq n$, where $i, n \in \mathbb{N} \cup \{0\}$. For example, assuming that MATLAB uses 64-bit doubles with 52 fraction bits (as specified by IEEE 754), the rounding error for the fraction $\frac{1}{3}$ is approximately

$$err \approx 1.01010101 \cdot 2^{-54} \approx 5.61 \cdot 10^{-17} \qquad (43)$$

However, once this relatively small error is multiplied by ten 27 times to produce y_{30}, it grows substantially to

$$err \cdot 10^{27} \approx 5.61 \cdot 10^{10} \qquad (44)$$

This is just the approximate accumulated rounding error for the error introduced when computing $\frac{1}{3}$ to compute y_3, but many other terms contribute similar error, which helps explain the enormous discrepancy between the theoretical and computed behavior of the sequence $y_0, y_1, y_2, ..., y_{30}$. In summary, the naive algorithm is *unstable*, as it leads to exponential accumulation of rounding error, making it ineffective for computing y_n accurately.

(c) Rearranging the formula from the naive algorithm to express y_{n-1} in terms of y_n, we find

$$y_{n-1} = \frac{y_n - \frac{1}{n}}{-10} = \frac{\frac{1}{n} - y_n}{10} \qquad (45)$$

Thus, we propose the following algorithm to compute the sequence $y_0, y_1, ..., y_N$ for some $N \in \mathbb{N}$:

(a) Pick some (large) $M \geq n$ and set $y_M = 0$.

(b) Recursively compute $y_{n-1} = \frac{1}{10}(\frac{1}{n} - y_n)$ for all $n \in \{1, ..., n\}$.

(d) Define a_i to be the absolute computational error involved in computing y_i with the algorithm proposed in (c). For a given ε and a fixed N, we want to find an M such that

$$a_n < \varepsilon \text{ for all } n \in \{0, 1, ..., N\} \tag{46}$$

Note that $a_M = y_M$ since the computed (assumed) value of y_M is 0.
Claim: for all $0 \leq n \leq M$, $n \in \mathbb{N}$

$$a_n = a_M(\frac{1}{10})^{M-n} \tag{47}$$

Proof. We induct on n.
Base Case: $n = M$, we have

$$a_n = a_M = a_M(\frac{1}{10})^{M-n} = a_M(\frac{1}{10})^0 \tag{48}$$

so (47) holds for the base case.
Inductive Hypothesis: Assume (47) holds for all $0 < k \leq n \leq M$.
Inductive Step: Consider $n = k - 1$. Using the algorithm from (c), we compute

$$y_n = y_{k-1} = \frac{1}{10}(\frac{1}{k} - y_k) = \frac{1}{10 \cdot k} - \frac{y_k}{10} \tag{49}$$

Assuming calculations are performed with infinite precision, there is no additional error introduced by the $\frac{1}{10 \cdot k}$ computation, so all the error in $a_n = a_{k-1}$ must come from the computation of $-\frac{y_k}{10}$. By the inductive hypothesis, we know the computational error for computing y_k is $a_k = a_M(\frac{1}{10})^{M-k}$. Thus, we have

$$a_n = a_{k-1} = |-\frac{a_k}{10}| = |-\frac{a_M(\frac{1}{10})^{M-k}}{10}| = |-a_M(\frac{1}{10})^{M-k+1}| = a_M(\frac{1}{10})^{M-(k-1)} = a_M(\frac{1}{10})^{M-n} \tag{50}$$

The conclusion that (47) holds follows by induction for all $0 \leq n \leq M$, $n \in \mathbb{N}$.
This result allows us to rewrite (46) as

$$\text{For all } \varepsilon > 0, n \in \mathbb{N} \cup \{0\}, \quad \exists M \in \mathbb{N} \text{ such that } a_M(\frac{1}{10})^{M-n} < \varepsilon \text{ for all } n \in \{0, 1, ..., N\} \tag{51}$$

From (a), we know that y_n monotonically decreases to 0 as $n \to \infty$, so the error $a_M = y_M$ also decreases monotonically to 0 as $M \to \infty$. This means that

$$y_0 = \sup\{a_M | M \in \mathbb{N} \cup \{0\}\} \tag{52}$$

which implies

$$a_M \leq y_0 \text{ for all } M \in \mathbb{N} \cup \{0\} \tag{53}$$

Also, we can directly evaluate that

$$y_0 = \int_0^1 \frac{1}{x + 10} dx \leq \int_0^1 \frac{1}{10} dx = \frac{1}{10} \tag{54}$$

since $\frac{1}{x+10}$ is monotonically decreasing in x for all $x \in (-10, \infty)$ (and thus for all $x \in (0, 1) \subseteq (-10, \infty)$), and the value of $\frac{1}{x+10}$ at $x = 0$ is $\frac{1}{10}$. Thus, it suffices to find a $M \in \mathbb{N}$ such that

$$a_n = a_M(\frac{1}{10})^{M-n} \leq \frac{1}{10}(\frac{1}{10})^{M-n} = (\frac{1}{10})^{M-n+1} < \varepsilon \tag{55}$$

Rearranging the inequality from the right hand side of (55) to isolate M, we find

$$\begin{aligned}(\frac{1}{10})^{M-n+1} < \varepsilon \iff & \quad (M-n+1)ln(\frac{1}{10}) < ln(\varepsilon) \\ \iff & \quad (M-n+1) > \frac{ln(\varepsilon)}{ln(\frac{1}{10})} \\ \iff & \quad M > (n-1) - \frac{ln(\varepsilon)}{ln(10)} \qquad (56)\end{aligned}$$

Note that $(n-1) - \frac{ln(\varepsilon)}{ln(10)}$ is linearly increasing in n, so picking

$$M \in \mathbb{N} \text{ such that } M > (N-1) - \frac{ln(\varepsilon)}{ln(10)} \qquad (57)$$

guarantees that

$$a_n < \varepsilon \text{ for all } n \in \{0, 1, ..., N\} \qquad (58)$$

That is, assuming infinite precision calculations, picking an $M \in \mathbb{N}$ as described in (57) guarantees that the absolute error in the calculations of $y_0, y_1, .., y_N$ will each be less than ε.

(e) Rounding errors in the computer do not produce excessive errors using this algorithm because the rounding error exponentially decreases, as opposed to the exponential growth of the rounding error with the naive algorithm. With the algorithm from (c), we have

$$y_{n-1} = \frac{1}{10}(\frac{1}{n} - y_n) \qquad (58)$$

so both the rounding error from $\frac{1}{n}$ and the total error from computing y_n are divided by 10 when computing y_{n-1}. Let $b_i :=$ the rounding error introduced by computing $\frac{1}{i}$, for all $i \in \mathbb{N}$. Then, by the time y_3 is computed with the algorithm from (c), the rounding error from computing $\frac{1}{30}$ only contributes

$$b_{30} \cdot (\frac{1}{10})^{27} = b_{30} \cdot 10^{-27} < 10^{-27} \qquad (59)$$

with the last inequality following because the rounding error for $\frac{1}{30}$ is trivially less than 1. Comparing (59) with (44), we see that the exponential decay of rounding error with the algorithm from (c) means the computer does not produce excessive errors using this algorithm, whereas the exponential growth of rounding error with the naive algorithm led to enormous absolute and relative errors in computations.

(f) First, we must pick a value for $M \in \mathbb{N}$ which will guarantee that our algorithm finds the value of y_{30} to the specified accuracy. Plugging $\varepsilon = 10^{-5}$ and $N = 30$ into (57), we find that we will need an

$$M > (30-1) - \frac{ln(10^-5)}{ln(10)} = 29 - \frac{-5ln(10)}{ln(10)} = 29 + 5 = 34 \qquad (58)$$

The result from (58) suggests that $M = 35$ should be sufficient to compute y_{30} within 10^{-5} of its exact value. However, the inequality from (57) assumes infinitely precise calculations, which are impossible in practice. Thus, we choose to add 5 to this theoretical result and pick $M = 40$ to increase confidence that the error in computing y_{30} does not exceed 10^{-5}. We use the following MATLAB code:

```
y40 = 0;
y = zeros(39);
y(39) = (1/10)*(1/40)
i = 38;
while i > 0
    y(i) = (1/10)*(1/(i+1)-y(i+1));
    i=i-1;
end
```

to compute that

$$y_{30} \approx 0.002940928704639 \tag{59}$$

with accuracy of 10^{-5}. That is, the algorithm from (c) allows us to compute that

$$y_{30} \in [0.0029309, 0.0029509] \tag{60}$$

which completes the problem.

Problem 5

Supposed $f \in C^3$ (this means it has three derivatives and the third order derivative is continuous). We showed in class that the discretization error when using the difference quotient

$$\frac{f(a+h) - f(a)}{h} \tag{1}$$

as an approximation for $f'(a)$ is $O(h)$. Another difference quotient that can be used to approximate $f'(a)$ is

$$\frac{f(a+h) - f(a-h)}{2h} \tag{2}$$

You may have noticed from graphs that (2) tends to be more accurate than (1). Show that the discretization error when using (2) is $\mathcal{O}(h^2)$. If $h \in [-1, 1]$, what is the constant C so that

$$\left| \frac{f(a+h) - f(a-h)}{2h} - f'(a) \right| \leq Ch^2$$

Solution

To show the discretization error when using (2) is $\mathcal{O}(h^2)$, we use the Taylor Remainder Theorem. Since f can has three derivatives, and $f^{'''}(x)$ is continuous, we have

$$f(a+h) = f(a) + hf'(a) + \frac{h^2}{2!}f^{''}(a) + \frac{h^3}{3!}f^{'''}(b_1) \tag{61}$$

for some $b_1 \in [a, a+h]$ and

$$f(a+h) = f(a) - hf'(a) + \frac{h^2}{2!}f^{''}(a) - \frac{h^3}{3!}f^{'''}(b_2) \tag{62}$$

for some $b_2 \in [a-h, a]$. Subtracting (62) from (61) yields

$$f(a+h) - f(a-h) = 2hf'(a) + \frac{h^3}{6}(f^{'''}(b_1) + f^{'''}(b_2)) \tag{63}$$

Dividing both sides of (63) by $2h$, subtracting $f'(a)$, and taking the absolute value yields

$$\left| \frac{f(a+h) - f(a-h)}{2h} - f'(a) \right| = \frac{h^2}{12}|(f^{'''}(b_1) + f^{'''}(b_2))| \tag{64}$$

Since $f^{'''}$ is continuous, it is bounded, so

$$\exists M_1 > 0, \in \mathbb{R} \text{ such that } |f^{'''}(x)| \leq M_1 \text{ for all } x \in [a, a+h] \tag{65}$$

and

$$\exists M_2 > 0, \in \mathbb{R} \text{ such that } |f^{'''}(x)| \leq M_2 \text{ for all } x \in [a-h, a] \tag{66}$$

If we define $M := M_1 + M_2$, then (65) and (66) combine to imply

$$\left|\frac{f(a+h) - f(a-h)}{2h} - f'(a)\right| \leq h^2 \frac{M}{12} \tag{67}$$

for all $h \in [-1, 1]$, which implies that

$$\left|\frac{f(a+h) - f(a-h)}{2h} - f'(a)\right| = \mathcal{O}(h^2) \tag{68}$$

From (67), we also find that the constant C which guarantees that

$$\left|\frac{f(a+h) - f(a-h)}{2h} - f'(a)\right| \leq Ch^2$$

holds is $C = \frac{M}{12}$, where M is the sum of the supremum of $f'''(x)$ over $[a, a+h]$ and the supremum of $f'''(x)$ over $[a-h, a]$.

Assignment 2

Problem 1

(a) Suppose the binary (base two) representation of a number x is $1.1011 \times \text{two}^{110}$. What is the base ten representation of the number?

(b) How is fifty-seven fourths represented in base four?

Solution

(a) First, we convert the exponent to decimal to find

$$x = 1.1011 \cdot \text{two}^{0+2+4} = 1.1011 \cdot \text{two}^{6}$$

where subscripts denote base. Noting that $2^6 = 64$ in decimal, we find

$$x = 64 + 32 + 0(16) + 8 + 4 = 108$$

where subscripts denote base.

(b) In decimal, we can write $\frac{57}{4}$ as the following sum:

$$\frac{57}{4} = \frac{56}{4} + \frac{1}{4} = 14 + \frac{1}{4} = 3(4) + 2(1) + \frac{1}{4} = 3 \cdot 4^1 + 2 \cdot 4^0 + 1 \cdot 4^{-1}$$

Thus, we can write $\frac{57}{4}$ in base four as

$$\frac{57}{4} = 3.21 \cdot \text{four}^{1} = 3.21 \cdot \text{four}$$

Problem 2

Consider a computer where $(\beta, t, L, U) = (4, 5 - 2, 2)$.

(a) Which of the following numbers are floating point numbers in this computer? Select all that apply and explain briefly.

(i) one sixty-fourth

(ii) three sixteenths

(iii) one third

(iv) seven

(v) two hundred and fifty-six

(b) (i) What is the largest floating point number in this computer?

(ii) What is the smallest positive floating point number?

(iii) How many floating point numbers are there total? (Remember 0 but ignore "non-normal" numbers.)

(iv) What is the value of the rounding unit η.

(c) What is the distance between the smallest positive floating point number and the next smallest floating point number?

(d) What is the distance between the largest floating point number and the second largest floating point number?

Solution

(a) Only three sixteenths and seven are floating point numbers in this computer.

(i) Note that

$$\frac{1}{64} = \frac{1}{4^3} = 4^{-3}$$

That is, to satisfy the floating point requirement for a normalized number with first digit $d_0 > 0$, the decimal value $\frac{1}{64}$ would be expressed in our base $\beta = 4$ as

$$1.0 \cdot \text{four}^{-3}$$

But the lower bound on our exponent is $L = -2 > -3$, so this computer cannot produce any numbers with -3 as the exponent. Thus, one sixty-fourth is *not* a floating point number in this computer, as its exponent is outside $\{L, L+1, ..., U-1, U\} = \{-2, -1, 0, 1, 2\}$.

(ii) Note that

$$\frac{3}{16} = \frac{3}{4^2} = 3(4^{-2})$$

That is, to satisfy the floating point requirement for a normalized number with the first digit $d_0 > 0$, the decimal value $\frac{3}{16}$ would be expressed in our base $\beta = 4$ as

$$3.0 \cdot \text{four}^{-2}$$

Since our exponent $e = -2 \in \{L, L+1, \cdots, U-1, U\} = \{-2, -1, 0, 1, 2\}$ and we used less than $t = 5$ digits to represent the number, we know three sixteenths *is* a floating point number in our computer.

(iii) Note that

$$\frac{1}{3} = 0.\overline{33} > \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \frac{1}{256} + \frac{1}{1024} = 4^{-1} + 4^{-2} + 4^{-3} + 4^{-4} + 4^{-5} \approx 0.3330$$

In base $\beta = 4$, we can write this approximation as

$$1.1111 \cdot \text{four}^{-1}$$

Note that this value uses all $t = 5$ digits available to this computer, but it still fails to produce the exact value of one third. Note also that

$$\frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \frac{1}{256} + \frac{2}{1024} \approx 0.33398 > \frac{1}{3}$$

Thus, one third cannot be written exactly as a sequence of 5 digits in base four, so one this is *not* a floating point number in this computer.

(iv) Note that

$$7 = 4 + 3 \cdot 1 = 4^1 + 3 \cdot 4^0$$

Thus, we can write the number 7 in base $\beta = 4$ as

$$1.3 \cdot \text{four}^1$$

Since the exponent $e \in \{L, L+1, ..., U-1, U\} = \{-2, -1, 0, 1, 2\}$ and we used less than $t = 5$ digits to represent the number, we know seven *is* a floating point number in this computer.

(v) Note that

$$256 = 4^4$$

so it can be written in base $\beta = 4$ as

$$1.0 * \text{four}^4$$

However, $e = 4 \notin \{L, L+1, ..., U-1, U\}$. Thus, we know two hundred and fifty-six is *not* a floating point number in this computer

(b) (i) The largest floating point number in this computer is

$$3.3333 \cdot \text{four}^2$$

In decimal (base 10), this can be written as

$$3(4^2)+3(4^1)+3(4^0)+3(4^{-1})+3(4^{-2}) = 3(16)+3(4)+3(1)+3(\frac{1}{4})+3(\frac{1}{16}) = 48+12+3+\frac{12}{16}+\frac{3}{16} = \frac{1023}{16} = 63.9375$$

Note that this value is $\frac{1}{16}$ less than $4^3 = 4^{U+1} = 64$

(ii) The smallest positive floating point number in this computer is

$$1.0000 \cdot \text{four}^L = 1.0000 \cdot \text{four}^{-2}$$

In base 10, this can be written as

$$1(4^{-2}) = \frac{1}{4^2} = \frac{1}{16} = 0.0625$$

(iii) Every nonzero floating point number in this computer is of the form

$$x = \pm d_0.d_1d_2d_3d_4 \cdot \text{four}^e$$

where $d_0 \in \{1,2,3\}$, $d_1, d_2, d_3, d_4 \in \{0,1,2,3,4\}$, and $e \in \{L, ..., U\} = \{-2,-1,0,1,2\}$. Thus, we have 2 choices for the sign, 3 choices for d_0, 4 choices for each of d_1, d_2, d_3, d_4, and 5 choices for e. This yields

$$2 \cdot 3 \cdot 4 \cdot 5 = 5! = 120$$

different possible numbers. However, we also have to consider 0. Thus, our computer can produce

$$120 + 1 = 121$$

floating point numbers, including 0 but ignoring "non-normal" numbers.

(iv) By definition, the rounding unit is

$$\eta = \frac{1}{2}\beta^{-(t-1)} = \frac{1}{2}\beta^{-(5-1)} = \frac{1}{2}\beta^{-4} = \frac{1}{2\beta^4} = \frac{1}{2(4)^4} = \frac{1}{2(256)} = \frac{1}{512} = 0.001953125$$

(c) We found in part b.ii that the smallest positive floating point number in this computer is

$$\frac{1}{16} = 0.0625$$

which has the form

$$1.0000 \cdot \text{four}^{-2}$$

in base $\beta = 4$. Thus, the next smallest positive floating point number is

$$1.0001 \cdot \text{four}^{-2}$$

In base 10, we can write this as

$$(1 + 0(\frac{1}{4}) + 0(\frac{1}{16}) + 0(\frac{1}{64}) + 1(\frac{1}{256}))\frac{1}{16} = (\frac{257}{256})\frac{1}{16} = \frac{257}{4096} \approx 0.062744$$

Subtracting the smallest positive floating point number from the next smallest, we find the distance between the two is

$$(\frac{257}{256})\frac{1}{16} - \frac{1}{16} = \frac{1}{16}(\frac{257}{256} - 1) = \frac{1}{16}(\frac{257}{256} - \frac{256}{256}) = \frac{1}{16}(\frac{1}{256}) = \frac{1}{4096} \approx 2.4414 \cdot 10^{-4}$$

(d) In part b.i, we found that the largest possible floating point number in this computer is

$$63.9375 = \frac{1023}{16}$$

which can be written as a normalized number in base $\beta = 4$ as

$$3.3333 \cdot \text{four}^2$$

Thus, the second largest floating point number has the form

$$3.3332 \cdot \text{four}^2$$

In base 10, we can write this as

$$3(4^2) + 3(4^1) + 3(4^0) + 3(4^{-1}) + 2(4^{-2}) = 3(16) + 3(4) + 3 + \frac{3}{4} + \frac{2}{16} = \frac{1022}{16} = 63.875$$

Subtracting the second largest positive floating point number from the largest positive floating point number, we find the distance between the two is

$$\frac{1023}{16} - \frac{1022}{16} = \frac{1023 - 1022}{16} = \frac{1}{16}$$

Note that this distance is $4^4 = 256$ times larger than the distance between the two smallest positive floating point numbers. This makes sense, as we have a fixed number $t = 5$ of digits, so the precision decrease by 4^4 as the exponent of the first digit d_0 increases by 4.

Problem 3

(a) Provide an example using $(\beta, t, L, U) = (10, 4, -2, 3)$ to illustrate that if x and y are floating point numbers then xy need not be a floating point number. Do this without having x, y, or xy be larger than the largest floating point number of smaller than the smallest floating point number. Explain briefly.

(b) Suppose x and $y \neq 0$ are real numbers. Find a bound on the relative error when the product xy is calculated in the computer as $fl(x) \times_{\text{algorithm}} fl(y))$.

Solution

(a) Let

$$x = 5.001 \cdot 10^3 = 5001$$

and

$$y = 5.000 \cdot 10^{-1} = \frac{5}{10} = \frac{1}{2}$$

Then

$$x \cdot y = \frac{5001}{2} = 2500.5 = 2.5005 \cdot 10^3$$

Thus, the exact value of xy takes $5 > t = 4$ digits to write in base $\beta = 10$, so

$$xy \neq d_0.d_1d_2d_3 \cdot 10^e$$

for all $d_0 \in \{1, ..., 9\}$, $d_1, d_2, d_3 \in \{0, ..., 9\}$, and $e \in \{L, ..., U\} = \{-2, ..., 3\}$. Thus xy is not a floating point number for a computer with $(\beta, t, L, U) = (10, 4, -2, 3)$. Also, note that

$$1.000 \cdot 10^{-2} < 5.001 \cdot 10^3 < 9.999 \cdot 10^3,$$
$$1.000 \cdot 10^{-2} < 5.000 \cdot 10^{-1} < 9.999 \cdot 10^3, \quad \text{and}$$
$$1.000 \cdot 10^{-2} < 2500.5 < 9.999 \cdot 10^3$$

so x, y, and xy are all between the smallest positive floating point number and the largest positive floating point number. Thus, $x = 5001$, $y = \frac{1}{2}$, $xy = 2500.5$ combine to illustrate that xy might *not* be a floating point number, even if x and y both are.

(b) We want to compute the relative error

$$\left|\frac{x \cdot y - fl(fl(x) \cdot_{\text{algorithm}} fl(y))}{x \cdot y}\right|$$

From lecture, we know that, for any real number x, we can write

$$fl(x) = x(1+\varepsilon)$$

for some ε such that $|\varepsilon| \leq \eta$, where $\eta := \frac{1}{2}\beta^{-(t-1)}$ is the rounding unit. Thus, we can write

$$fl(x) = x(1+\varepsilon_1)$$

for some $|\varepsilon_1| \leq \eta$ and

$$fl(y) = y(1+\varepsilon_2)$$

for some $|\varepsilon_2| \leq \eta$. This allows us to rewrite the relative error of our multiplication computation as

$$\left|\frac{x \cdot y - fl(fl(x) \cdot_{\text{algorithm}} fl(y))}{x \cdot y}\right| = \left|\frac{xy - fl(x(1+\varepsilon_1) \cdot_{\text{algorithm}} y(1+\varepsilon_2))}{xy}\right|$$

We can apply the same identity once more to find

$$fl(x(1+\varepsilon_1) \cdot_{\text{algorithm}} y(1+\varepsilon_2)) = x(1+\varepsilon_1)y(1+\varepsilon_2)(1+\varepsilon_3) = xy(1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3)$$

for some $|\varepsilon_3| \leq \eta$. Plugging this into the relative error equation yields

$$\begin{aligned}
\left|\frac{x \cdot y - fl(fl(x) \cdot_{\text{algorithm}} fl(y))}{x \cdot y}\right| &= \left|\frac{xy - xy(1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3)}{xy}\right| \\
&= \left|\frac{xy}{xy}(1 - (1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3))\right| \\
&= \left|(1 - (1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3))\right|
\end{aligned}$$

where the last equality follows because $x, y \neq 0 \implies xy \neq 0$. Expanding, we find

$$\begin{aligned}
\left|\frac{x \cdot y - fl(fl(x) \cdot_{\text{algorithm}} fl(y))}{x \cdot y}\right| &= |1 - (1 + \varepsilon_1 + \varepsilon_2 + \varepsilon_1\varepsilon_2)(1+\varepsilon_3)| \\
&= |1 - (1 + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_1\varepsilon_2 + \varepsilon_1\varepsilon_3 + \varepsilon_2\varepsilon_3 + \varepsilon_1\varepsilon_2\varepsilon_3)| \\
&= |-(\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_1\varepsilon_2 + \varepsilon_1\varepsilon_3 + \varepsilon_2\varepsilon_3 + \varepsilon_1\varepsilon_2\varepsilon_3)| \\
&= |\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_1\varepsilon_2 + \varepsilon_1\varepsilon_3 + \varepsilon_2\varepsilon_3 + \varepsilon_1\varepsilon_2\varepsilon_3|
\end{aligned}$$

Note that $|\varepsilon_1|, |\varepsilon_2|, |\varepsilon_3| \leq \eta$ combines with

$$\eta = \frac{1}{2}\beta^{-(t-1)} <= \frac{1}{2}1^{-(t-1)} = \frac{1}{2} < 1$$

to imply that

$$\varepsilon_1\varepsilon_2 < \varepsilon_1, \varepsilon_2 \qquad \varepsilon_1\varepsilon_3 < \varepsilon_1, \varepsilon_3, \quad \text{and} \quad \varepsilon_2\varepsilon_3 < \varepsilon_2, \varepsilon_3$$

Similarly,

$$\varepsilon_1\varepsilon_2\varepsilon_3 < \varepsilon_1\varepsilon_2, \varepsilon_2\varepsilon_3 \implies \varepsilon_1\varepsilon_2\varepsilon_3 < \varepsilon_1, \varepsilon_2, \varepsilon_3$$

Thus, we can group $\varepsilon_1\varepsilon_2 + \varepsilon_1\varepsilon_3 + \varepsilon_2\varepsilon_3 + \varepsilon_1\varepsilon_2\varepsilon_3 = O(\varepsilon^2)$ together as higher-order terms (h.o.t.) to find

$$\left|\frac{x \cdot y - fl(fl(x) \cdot_{\text{algorithm}} fl(y))}{x \cdot y}\right| = \left|\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \text{h.o.t.}\right|$$

Claim: $|x_1 + \cdots + x_n| \leq |x_1| + \cdots + |x_n|$ for all $n \in \mathbb{N}$, $x_1, ..., x_n \in \mathbb{R}$.
Proof. We induct on n.
Base Case: $n = 1$. Then $|x_1 + \cdots + x_n| = |x_1|$, so the claim is trivially true.
Inductive Hypothesis: Assume $|x_1 + \cdots + x_n| \leq |x_1| + \cdots + |x_n|$ for all $1 \leq n \leq k$, $n \in \mathbb{N}$, and all $x_1, ..., x_n \in \mathbb{R}$.
Inductive Step: Consider $n = k + 1$. Note that, if $x_1 + \cdots + x_k$ and x_{k+1} have the same sign, then

$$\left|x_1 + \cdots + x_k + x_{k+1}\right| = \left|x_1 + \cdots + x_k\right| + |x_{k+1}|$$

On the other hand, if $x_1 + \cdots + x_k$ and x_{k+1} have different signs, then

$$|x_1 + \cdots + x_k + x_{k+1}| < |x_1 + \cdots + x_k| + |x_{k+1}|$$

Thus, for all $x_1, ..., x_{k+1}$, we have

$$|x_1 + \cdots + x_k + x_{k+1}| \leq |x_1 + \cdots + x_k| + |x_{k+1}|$$

By the inductive hypothesis, we know

$$|x_1 + \cdots + x_k| \leq |x_1| + \cdots + |x_k|$$

Using this result, we find

$$|x_1 + \cdots + x_k + x_{k+1}| \leq |x_1| + \cdots + |x_k| + |x_{k+1}|$$

which is exactly what we want to show.
The conclusion that $|x_1 + \cdots + x_n| \leq |x_1| + \cdots + |x_n|$ follows by induction for all $n \in \mathbb{N}$ and all $x_1, ..., x_n \in \mathbb{R}$.
Applying this inequality to our equation for relative error yields

$$\left|\frac{x \cdot y - fl(fl(x) \cdot_{\text{algorithm}} fl(y))}{x \cdot y}\right| \leq |\varepsilon_1| + |\varepsilon_2| + |\varepsilon_3| + |\text{h.o.t.}|$$

By definition, $|\varepsilon_1|, |\varepsilon_2|, |\varepsilon_3| \leq \eta$. This allows us to conclude

$$\left|\frac{x \cdot y - fl(fl(x) \cdot_{\text{algorithm}} fl(y))}{x \cdot y}\right| \leq 3\eta + |\text{h.o.t.}| = \frac{3}{2}\beta^{-(t-1)} + |\text{h.o.t}|$$

where h.o.t. denotes the higher order terms. This is our upper bound for the relative error from computing the product of two real numbers $x, y \neq 0$. Note that this upper bound is the same as the one we found in lecture for the relative error of division.

Problem 4

Suppose a machine with a floating point system $(\beta, t, L, U) = (10, 8, -50, 50)$ is used to find the roots of the quadratic equation

$$ax^2 + bx + c = 0$$

using the standard formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Describe the numerical difficulties that arise in each of the following cases and, in each case, suggest a different way of calculating the roots that would be more accurate or explain why no such method exists.

(a) $a = 1$; $b = -10^5$; $c = 1$

(b) $a = 5 \cdot 10^{30}$; $b = 5 \cdot 10^{30}$; $c = -4 \cdot 10^{30}$

(c) $a = 10^{-30}$; $b = -10^{30}$; $c = 10^{30}$

Solution

(a) When $a = 1$, $b = -10^5$, and $c = 1$, the equation simplifies to

$$x = \frac{-b \pm \sqrt{b^2 - 4}}{2}$$

so our roots are

$$x_1 = \frac{1}{2}(-b + \sqrt{b^2 - 4}) = \frac{1}{2}(-(-10^5) + \sqrt{(-10^5)^2 - 4}) = \frac{1}{2}(10^4 + \sqrt{10^{10} - 4})$$

and

$$x_2 = \frac{1}{2}(-b - \sqrt{b^2 - 4}) = \frac{1}{2}(10^5 - \sqrt{10^{10} - 4})$$

Note that

$$10^{10} - 4 = 10000000000 - 4 = 9999999996 = \underbrace{9.999999996}_{\text{10 digits}} \cdot 10^9$$

Since there are $10 > t = 8$ digits in the exact result, the computer will calculate $10^{10} - 4$ by rounding based on whether the ninth digit satisfies $d_8 \geq \frac{\beta}{2} = \frac{10}{2} = 5$. Here, $d_8 = 9 \geq 5$, so the computer will round $9.999999996 \cdot 10^9$ up to

$$1.0000000 \cdot 10^{10} = b^2$$

That is, the computer will return

$$b^2 - 4 = b^2$$

For x_2, this yields to a computed value of

$$\hat{x_2} = \frac{1}{2}(10^5 - \sqrt{10^{10}}) = \frac{1}{2}(10^5 - 10^5) = 0$$

which yields to a relative error of

$$\left|\frac{x_2 - \hat{x_2}}{x_2}\right| = \left|\frac{x_2}{x_2}\right| = 1$$

or 100%. Thus, although $\hat{x_1}$ could be computed with better accuracy, there is no way to compute $\hat{x_2}$ with low relative error using the standard quadratic formula.

However, we can note that

$$x_1 \cdot x_2 = \frac{1}{4}(-b + \sqrt{b^2 - 4})(-b - \sqrt{b^2 - 4}) = \frac{1}{4}(b^2 - (b^2 - 4)) = \frac{1}{4}(4) = 1$$

where the second equality follows from the identity $(a - b)(a + b) = a^2 - b^2$. Dividing both sides by x_1 since $x_1 = \frac{1}{2}(10^4 + \sqrt{10^{10} - 4}) > 0$, we find

$$x_2 = \frac{1}{x_1}$$

Thus, we can compute $\hat{x_1}$ first. Note that computing $\hat{x_1}$ only involves addition of values with the same sign (and dividing by 2), and

$$x_1 = \frac{1}{2}(10^5 + \sqrt{10^{10} - 4}) \approx \frac{1}{2}(10^5 + 10^5) = 10^5 << 9.9999999 \cdot 10^{50}$$

Thus, we should not have any trouble computing $\hat{x_1}$ due to overflow nor subtracting close numbers, so we should compute $\hat{x_1}$ relatively accurately. From there, we can compute $\hat{x_2}$ using the identity $x_2 = \frac{1}{x_1}$. Note that

$$x_1 \approx 10^5 \implies x_2 \approx \frac{1}{10^5} = 10^{-5} >> 1.0000000 \cdot 10^{-50}$$

so we should be able to compute $\hat{x_2}$ from $\hat{x_1}$ relatively accurately without running into overflow problems. Thus computing $\hat{x_2}$ using $x_2 = \frac{1}{x_1}$ would be more accurate than using the standard quadratic formula in this case.

(b) When $a = 5 \cdot 10^{30}$, $b = 5 \cdot 10^{30}$, $c = -4 \cdot 10^{30}$, the equation simplifies to

$$x = \frac{-b \pm \sqrt{b^2 - 4(5 \cdot 10^{30})(-4 \cdot 10^{30})}}{2(5 \cdot 10^{30})} = \frac{-b \pm \sqrt{b^2 + 80 \cdot 10^{60}}}{10^{31}}$$

Note that

$$80 \cdot 10^{60} = 8.0 \cdot 10^{61} > 9.9999999 \cdot 10^{50}$$

where the value on the right hand side is the largest positive floating point number in this computer. Thus, the computer will struggle to compute either of the roots accurately because it will overflow when computing $-4ac$ for both roots.

However, we can note that a, b, and c each have a factor of 10^{30} in them, which leads $b^2 - 4ac$ to have a factor of $(10^{30})^2$. That is,

$$x = \frac{-b \pm \sqrt{(5 \cdot 10^{30})^2 + 80(10^{30})^2}}{10^{31}} = \frac{-b \pm \sqrt{(10^{30})^2(25 + 80)}}{10^{31}} = \frac{-b \pm 10^{30}\sqrt{105}}{10^{31}} = \frac{-5 \cdot 10^{30} \pm 10^{30}\sqrt{105}}{10^{31}}$$

so our two roots are

$$x_1 = \frac{-5 \cdot 10^{30} + 10^{30}\sqrt{105}}{10^{31}} = \frac{1}{10}\frac{10^{30}(\sqrt{105} - 5)}{10^{30}} = \frac{1}{10}(\sqrt{105} - 5)$$

and

$$x_2 = \frac{-5 \cdot 10^{30} - 10^{30}\sqrt{105}}{10^{31}} = \frac{1}{10}\frac{10^{30}(-\sqrt{105} - 5)}{10^{30}} = -\frac{1}{10}(\sqrt{105} + 5)$$

Note that we will have some error computing the irrational number $\sqrt{105}$ with anything less than infinite precision, but we should get a much more accurate answer than the standard formula yields since there will be no overflow error. Moreover, to be cautious, we could limit potential cancellation error from $\sqrt{105} - 5$ by using the alternative method from part (a) (using $x_1 x_2 = \frac{c}{a}$ instead of $x_1 x_2 = 1$). Thus, using $x_1 = \frac{1}{10}(\sqrt{105} - 5)$ and $x_2 = -\frac{1}{10}(\sqrt{105} + 5)$ would lead to more accurate computations than using the standard quadratic formula in this case.

(c) When $a = 10^{-30}$, $b = -10^{30}$, $c = 10^{30}$, the equation simplifies to

$$x = \frac{-b \pm \sqrt{b^2 - 4(10^{-30})(10^{30})}}{2 \cdot 10^{-30}} = \frac{-b \pm \sqrt{b^2 - 4}}{2 \cdot 10^{-30}}$$

so our roots are

$$x_1 = \frac{10^{30} + \sqrt{(10^{30})^2 - 4}}{2 \cdot 10^{-30}} = \frac{1}{2}10^{30}(10^{30} + \sqrt{(10^{30})^2 - 4}) = \frac{1}{2}(10^{60} + 10^{30}\sqrt{(10^{30})^2 - 4})$$

and

$$x_2 = \frac{10^{30} - \sqrt{(10^{30})^2 - 4}}{2 \cdot 10^{-30}} = \frac{1}{2}10^{30}(10^{30} - \sqrt{(10^{30})^2 - 4}) = \frac{1}{2}(10^{60} - 10^{30}\sqrt{(10^{30})^2 - 4})$$

Note that attempting to compute x_2 directly leads to overflow with the initial two calculations of 10^{60}, potential rounding error with the calculation of $10^{60}-4$, potential overflow from the computation of $10^{30}\cdot\sqrt{(10^{30})^2-4}$, and potential cancellation error like that seen in part (a) from subtracting

$$10^{30}\sqrt{(10^{30})^2-4}\approx 10^{30}\cdot 10^{30}=10^{60}$$

from 10^{60}. Thus, the directly computed value $\hat{x}_2$ is likely to be *very* inaccurate. Unfortunately,

$$x_1=\frac{1}{2}(10^{60}+10^{30}\sqrt{(10^{30})^2-4}\approx\frac{1}{2}(10^{60}+10^{30}\sqrt{10^{60}})=\frac{1}{2}(10^{60}+10^{30}\cdot 10^{30})=\frac{1}{2}(10^{60}+10^{60})=10^{60}>9.9999999\cdot 10^{50}$$

so this computer cannot even store a remotely accurate value of x_1. Attempting to directly compute x_1 would lead to all the same overflow error as computing x_2. However, there would be no cancellation error since the $\approx 10^{60}$ terms are added instead of subtracted. Although we cannot compute theoretically exact values of x_1 and x_2 using the formulae written above, we can approximate x_2. First, pull a factor of $10^{30}=\sqrt{10^{60}}$ out from the square root for both x_1 and x_2 to find

$$x_1=\frac{1}{2}(10^{60}+10^{60}\sqrt{1-4\cdot 10^{-60}})=\frac{10^{60}}{2}(1+\sqrt{1-4\cdot 10^{-60}})$$

and

$$x_2=\frac{1}{2}(10^{60}-10^{60}\sqrt{1-4\cdot 10^{-60}})=\frac{10^{60}}{2}(1-\sqrt{1-4\cdot 10^{-60}})$$

We still cannot directly compute the value inside the square root since 10^{-60} is smaller than the smallest positive floating point number in this system, and thus computing it will cause overflow. However, recall from discussion that the first order Taylor expansion of $\sqrt{1-x}$ at $a=0$ is

$$\sqrt{1-x}\approx f(a)+f'(a)(x-a)=\sqrt{1-a}-\frac{1}{2}(x-a)=\sqrt{1}-\frac{1}{2}x=1-\frac{x}{2}$$

Thus, we can replace $\sqrt{1-4\cdot 10^{-60}}$ with its first order Taylor approximation $1-2\cdot 10^{-60}$ to find

$$x_1\approx\frac{10^{60}}{2}(1+(1-2\cdot 10^{-60}))=\frac{10^{60}}{2}(2-2\cdot 10^{-60})=10^{60}(1-10^{-60})=10^{60}-1$$

and

$$x_2\approx\frac{10^{60}}{2}(1-(1-2\cdot 10^{-60}))=\frac{10^{60}}{2}(2\cdot 10^{-60})=1$$

This method allows us to approximate $x_2\approx 1$, a number we can easily store with this computer. However, we still cannot compute x_1 accurately, even with this approximation, as the approximate value of $x_1\approx 10^{60}-1$ is much larger than the largest floating point number this computer can store. Moreover, we cannot use the technique from (a) to solve for $x_1=\frac{c}{a}\frac{1}{x_2}$, as we cannot even store $\frac{c}{a}=\frac{10^{30}}{10^{-30}}=10^{60}$ accurately due to overflow. Thus, although we can utilize a first order Taylor polynomial to approximate x_2 accurately, we have no method of computing and storing x_1 without causing overflow. This suggests there is no way to accurately compute both roots in our computer for this case.

Problem 5

Suppose we estimate the derivative of a function f at a point $x=a$ in the computer using a difference quotient

$$f'(a)\approx\frac{f(a+h)-f(a)}{h}$$

There are three kinds of errors that are introduced as a result.

- A discretization error because $f'(x)$ is not exactly equal to $\frac{f(a+h)-f(a)}{h}$.
- Rounding errors in the calculation of $f(a+h)$ and $f(a)$.
- Rounding errors when computing $f(a+h)-f(a)$ and when computing $\frac{f(a+h)-f(a)}{f(h)}$.

The computed value of the difference quotient is then

$$(f_c(a+h) -_{\text{algorithm}} f_c(a)) \div_{\text{algorithm}} h$$

where f_c is the computed value of f (and is, therefore, necessarily a floating point number). Let's assume that h is a floating point number. Notice, the absolute error between the true value of $f'(a)$ and its computed value can be bounded by

$$\begin{aligned} |f'(a) - (f_c(a+h) -_{\text{algorithm}} f_c(a)) \div_{\text{algorithm}} h| & \\ \leq & \left|f'(a) - \frac{f(a+h)-f(a)}{h}\right| \\ + & \left|\frac{f(a+h)-f(a)}{h} - \frac{f_c(a+h)-f_c(a)}{h}\right| \\ + & \left|\frac{f_c(a+h)-f_c(a)}{h} - (f_c(a+h) -_{\text{algorithm}} f_c(a)) \div_{\text{algorithm}} h\right| \end{aligned}$$

We have already seen that the discretization error

$$\left|f'(a) - \frac{f(a+h)-f(a)}{h}\right| \leq \frac{Mh}{2}$$

where M is a bound on $|f"|$. Notice, furthermore, that, although we have written an inequality, the actual error is actually approximately equal to the right-hand side when h is small and M is replaced by $|f"(a)|$.

(a) Suppose the relative error when calculating the value of f is bounded by ε; in other words, for all real numbers x,

$$\frac{|f(x)-f_c(x)|}{|f(x)|} \leq \varepsilon$$

Show that

$$\left|\frac{f(a+h)-f(a)}{h} - \frac{f_c(a+h)-f_c(a)}{h}\right| \leq \frac{2K\varepsilon}{h}$$

where K is a bound on $|f|$. Notice, in this case, the actual error is just bounded by the right-hand-side and we don't expect it to be equal to what we have on the right; indeed, we would expect it to look somewhat random.

(b) Show that for small enough h

$$\begin{aligned} \left|\frac{f_c(a+h)-f_c(a)}{h} - (f_c(a+h) -_{\text{algorithm}} f_c(a)) \div_{\text{algorithm}} h\right| & \\ \leq & \quad 2\Big(\frac{2K\varepsilon}{h} + \frac{Mh}{2} + L\Big)\eta \end{aligned}$$

where L is a bound on f' and η is the rounding unit.

(d) Suppose $f(x) = sin(x)$ and $a = 1.2$. In this case, $f'(x) = cos(x)$ and it is reasonable to assume that $\varepsilon \approx \eta$. What are the values of M, K, and L? What is the value of h that minimizes the error bound?

(e) Use the computer to sketch the graph of the error between the actual value of $f'(x)$ and its computed value as a function of h. Use log scales on both axes with h ranging from about 10^{-20} to $10^0 = 1$. On the same axes, plot the discretization error $\frac{cos(1.2)h}{2}$ (Your picture should be similar to Figure 1.3 in Chen and Greif and the one we created in class). Provide your graph when you submit your homework.

(f) Explain your graph in (e) in light of parts (a)- (d). In particular, for what values of h is the error dominated by the discretization error and for what values is it dominated by the rounding error? Does this agree with your calculation in part (d)? Why does the graph level off when h is very small? At what value does that happen and why?

Solution

(a) Note that

$$\begin{aligned}\Big|\frac{f(a+h)-f(a)}{h}-\frac{f_c(a+h)-f_c(a)}{h}\Big| &= \Big|\frac{f(a+h)-f(a)-f_c(a+h)+f_c(a)}{h}\Big| \\ &= \Big|\frac{f(a+h)-f_c(a+h)}{h}-\frac{f(a)-f_c(a)}{h}\Big|\end{aligned}$$

By our proof in part (b) of **Problem 3**, we find

$$\begin{aligned}\Big|\frac{f(a+h)-f(a)}{h}-\frac{f_c(a+h)-f_c(a)}{h}\Big| &\leq \Big|\frac{f(a+h)-f_c(a+h)}{h}\Big|+\Big|-\frac{f(a)-f_c(a)}{h}\Big| \\ &= \Big|\frac{f(a+h)-f_c(a+h)}{h}\Big|+\Big|\frac{f(a)-f_c(a)}{h}\Big| \\ &= \frac{|f(a+h)-f_c(a+h)|}{|h|}+\frac{|f(a)-f_c(a)|}{|h|}\end{aligned}$$

Plugging in our assumption that

$$\frac{|f(x)-f_c(x)|}{|f(x)|}\leq\varepsilon \implies |f(x)-f_c(x)|\leq\varepsilon|f(x)|$$

for all $x\in\mathbb{R}$, we find

$$\Big|\frac{f(a+h)-f(a)}{h}-\frac{f_c(a+h)-f_c(a)}{h}\Big|\leq\frac{\varepsilon|f(a+h)|}{|h|}+\frac{\varepsilon|f(a)|}{|h|}$$

Applying the given assumption that $\exists K\in\mathbb{R}$ such that $|f(x)|\leq K$ for all $x\in\mathbb{R}$ (or at least on $[a,a+h]$), we find

$$\Big|\frac{f(a+h)-f(a)}{h}-\frac{f_c(a+h)-f_c(a)}{h}\Big|\leq\frac{1}{|h|}(\varepsilon K+\varepsilon K)=\frac{2K\varepsilon}{|h|}=\frac{2K\varepsilon}{h}$$

where the last equality assumes the step size $h>0$. This completes the proof that

$$\Big|\frac{f(a+h)-f(a)}{h}-\frac{f_c(a+h)-f_c(a)}{h}\Big|\leq\frac{2K\varepsilon}{h}$$

where ε is the bound on $\frac{|f(x)-f_c(x)|}{|f(x)|}$ for all $x\in\mathbb{R}$ and K is a bound on $|f|$.

(b) Note that

$$\begin{aligned}(f_c(a+h)-_{\text{algorithm}}f_c(a))\div_{\text{algorithm}}h &= fl(\frac{fl(f_c(a+h)-f_c(a))}{h})=fl(\frac{(f_c(a+h)-f_c(a))(1+\varepsilon_1)}{h}) \\ &= \frac{f_c(a+h)-f_c(a)}{h}(1+\varepsilon_1)(1+\varepsilon_2)\end{aligned}$$

for some $|\varepsilon_1|, |\varepsilon_2| \leq \eta$, which directly implies

$$\begin{aligned}
&\left|\frac{f_c(a+h)-f_c(a)}{h} - (f_c(a+h) -_{\text{algorithm}} f_c(a)) \div_{\text{algorithm}} h\right| \\
&= \left|\frac{f_c(a+h)-f_c(a)}{h} - \frac{f_c(a+h)-f_c(a)}{h}(1+\varepsilon_1)(1+\varepsilon_2)\right| \\
&= \left|\frac{f_c(a+h)-f_c(a)}{h}(1-(1+\varepsilon_1)(1+\varepsilon_2)\right| \\
&= \left|\frac{f_c(a+h)-f_c(a)}{h}(\varepsilon_1+\varepsilon_2+\text{h.o.t})\right| \\
&\leq \left|\frac{f_c(a+h)-f_c(a)}{h}\right|(\eta+\eta+|\text{h.o.t.}|) \\
&= \left|\frac{f_c(a+h)-f_c(a)}{h}\right|(2\eta+|\text{h.o.t.}|)
\end{aligned}$$

where the inequality follows by the triangle inequality and the upper bound on $|\varepsilon_1|$ and $|\varepsilon_2|$.

Claim: We claim that $|x| - |y| \leq |x-y|$ for all $x, y \in \mathbb{R}$.
Proof. We consider four cases based on the signs of x and y.

1. If $x > 0$ and $y > 0$, then
$$|x| = x, |y| = y \implies |x| - |y| = x - y \leq |x-y|$$
where the inequality follows since $a \leq |a|$ for all $a \in \mathbb{R}$ by definition of the absolute value function.
2. If $x > 0$ and $y \leq 0$, then
$$|x| = x, y = -|y| \implies |x| - |y| = x + y \leq x \leq x - y \leq |x-y|$$
where the first two inequalities follow from $y \leq 0$ and the final inequality follows from the definition of absolute value.
3. If $x \leq 0$ and $y \leq 0$, then
$$x = -|x|, y = -|y| \implies |x| - |y| = y - x \leq |y-x| = |-(y-x)| = |x-y|$$
where the inequality and the final two equalities follow by the definition of absolute value.
4. Finally, if $x \leq 0$ and $y > 0$, then
$$x = -|x|, y = |y| \implies |x| - |y| = -x - y \leq -x \leq y - x \leq |y-x| \leq |x-y|$$
where the first two inequalities follow from $y > 0$ and the last two follow by the definition of absolute value.

This completes the proof that $|x| - |y| \leq |x-y|$ for all $x, y \in \mathbb{R}$. Applying this result, we find

$$\left|\frac{f_c(a+h)-f_c(a)}{h}\right| - |f'(a)| \leq \left|\frac{f_c(a+h)-f_c(a)}{h} - f'(a)\right| = \left|f'(a) - \frac{f_c(a+h)-f_c(a)}{h}\right|$$

Applying the triangle inequality ($|x-z| \leq |x-y| + |y-z|$) and plugging in known upper bounds yields

$$\begin{aligned}
\left|\frac{f_c(a+h)-f_c(a)}{h}\right| - |f'(a)| \leq &\left|f'(a) - \frac{f(a+h)-f(a)}{h}\right| + \left|\frac{f(a+h)-f(a)}{h} - \frac{f_c(a+h)-f_c(a)}{h}\right| \\
\leq &\frac{Mh}{2} + \frac{2K\varepsilon}{h}
\end{aligned}$$

Adding $|f'(a)|$ to both sides and applying the given bound $|f'(x)| \leq L$, we find

$$\left|\frac{f_c(a+h) - f_c(a)}{h}\right| \leq \frac{Mh}{2} + \frac{2K\varepsilon}{h} + |f'(a)| \leq \frac{Mh}{2} + \frac{2K\varepsilon}{h} + L$$

Plugging this result into our inequality for $\left|\frac{f_c(a+h)-f_c(a)}{h} - (f_c(a+h) -_{\text{algorithm}} f_c(a)) \div_{\text{algorithm}} h\right|$ yields

$$\begin{aligned}\left|\frac{f_c(a+h) - f_c(a)}{h} - (f_c(a+h) -_{\text{algorithm}} f_c(a)) \div_{\text{algorithm}} h\right| & \\ \leq & (\frac{Mh}{2} + \frac{2K\varepsilon}{h} + L)(2\eta + |\text{h.o.t.}| \\ = & 2(\frac{2K\varepsilon}{h} + \frac{Mh}{2} + L)\eta + |\text{h.o.t}|(\frac{Mh}{2} + \frac{2K\varepsilon}{h} + L)\end{aligned}$$

For sufficiently small h, $f(h) = O(h^2) << g(h) = O(h)$, so the higher order term is negligible. This completes the proof that, for small enough h,

$$\begin{aligned}\left|\frac{f_c(a+h) - f_c(a)}{h} - (f_c(a+h) -_{\text{algorithm}} f_c(a)) \div_{\text{algorithm}} h\right| & \\ \leq & 2(\frac{2K\varepsilon}{h} + \frac{Mh}{2} + L)\eta\end{aligned}$$

where L is a bound on f' and η is the rounding unit.

(d) By definition, M is a bound on $|f"|$, K is a bound on $|f|$, and L is a bound on $|f'|$. If $f(x) = sin(x)$, then we know

$$-1 \leq f(x) = sin(x) \leq 1 \implies |f(x)| = |sin(x)| \leq 1$$

for all $x \in \mathbb{R}$, so we can take $K = 1$ to satisfy $|f(x)| \leq K$ for all $x \in \mathbb{R}$. We also have

$$-1 \leq f'(x) = \frac{d}{dx}sin(x) = cos(x) \leq 1 \implies |f'(x)| = |cos(x)| \leq 1$$

for all $x \in \mathbb{R}$, so we can take $L = 1$ to guarantee $|f'(x)| \leq L$ for all $x \in \mathbb{R}$. Finally, we have

$$-1 \leq f"(x) = \frac{d}{dx}cos(x) = -sin(x) \leq 1 \implies |f"(x)| = |-sin(x)| = |sin(x)| \leq 1$$

for all $x \in \mathbb{R}$, so we can take $M = 1$ to ensure $|f"(x)| \leq M$ for all $x \in \mathbb{R}$. Thus, the values of M, K, and L in this case are $M = K = L = 1$.
Plugging in $M = K = L = 1$ into the bound from part (b), we find

$$\begin{aligned}\left|\frac{f_c(a+h) - f_c(a)}{h} - (f_c(a+h) -_{\text{algorithm}} f_c(a)) \div_{\text{algorithm}} h\right| & \\ \leq & 2(\frac{2\varepsilon}{h} + \frac{h}{2} + 1)\eta \\ \approx & 2\eta(\frac{2\eta}{h} + \frac{h}{2} + 1)\end{aligned}$$

Note that

$$\frac{d}{dh}(\frac{2\eta}{h} + \frac{h}{2} + 1) = \frac{-2\eta}{h^2} + \frac{1}{2} = 0 \iff h^2 = 4\eta \iff h = 2(\pm\sqrt{\eta})$$

and

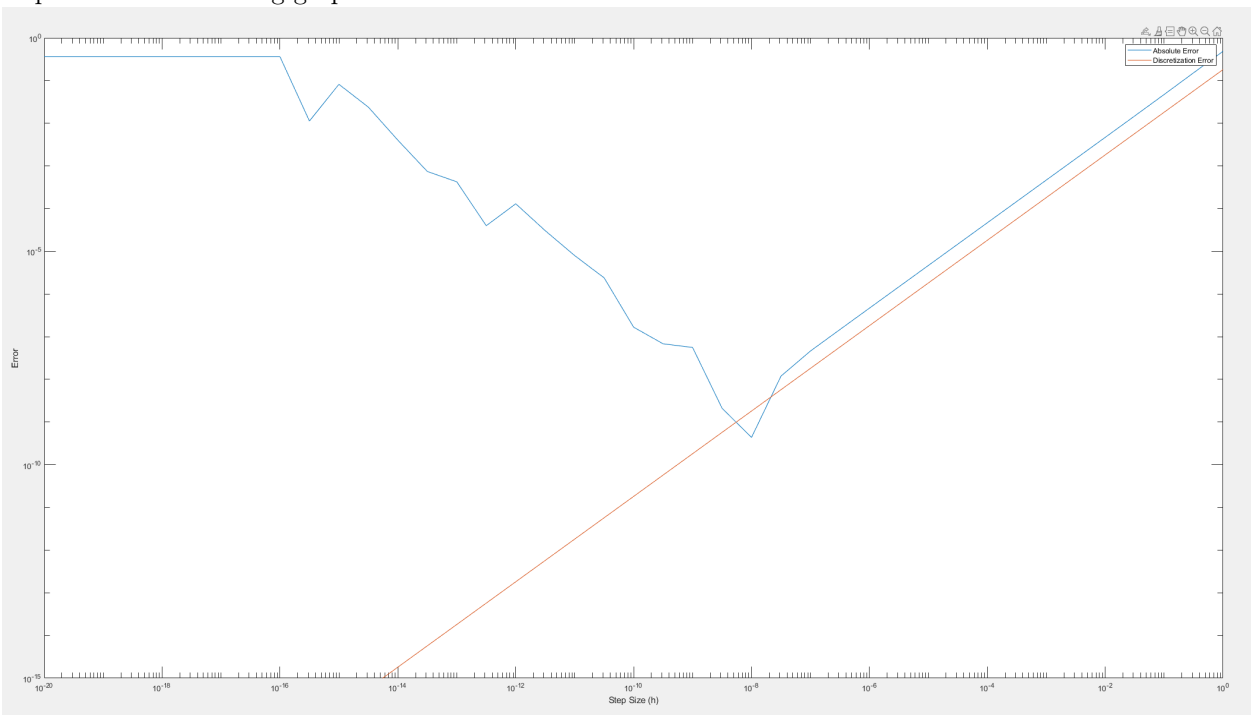
$$\frac{d^2}{dh^2} = \frac{d}{dh}(\frac{-2\eta}{h^2} + \frac{1}{2}) = \frac{4\eta}{h^3} > 0$$

for all $h > 0$. Thus, $\frac{2\eta}{h} + \frac{h}{2} + 1$ is decreasing in h for all $h \in (0, 2\sqrt{\eta})$ and increasing in h for all $h \in (2\sqrt{\eta}, \infty)$, so $h = 2\sqrt{n}$ is the unique $h > 0$ that minimizes $\frac{2\eta}{h} + \frac{h}{2} + 1$ (and thus $2\eta(\frac{2\eta}{h} + \frac{h}{2} + 1)$). Thus, the value of h that minimizes the error bound in this case is $h = 2\sqrt{\eta}$.

(e) We use the following MATLAB code:

```
a = 1.2;
i = -20:0.5:0;
h = 10.^i;
approx = (sin(a+h)-sin(a))./h;
val = cos(a);
abs_err = abs(val - approx);
disc_err = abs((cos(a).*h)./2);
loglog(h, abs_err);
hold on;
loglog(h, disc_err);
axis([10^(-20) 10^0 10^(-15) 10^0])
legend(["Absolute Error", "Discretization Error"])
xlabel("Step Size (h)");
ylabel("Error");
hold off;
```

to produce the following graph:



The discretization error is in red while the absolute error is in blue.

(f) The graph from (e) aligns with the expectations from the analysis in parts (a)-(d). The absolute error is dominated by the discretization error for larger values of h, specifically those $h >> 10^{-8}$. On the other hand, the absolute error is dominated by the rounding error for smaller values of h, specifically $h << 10^{-8}$. In parts (a) and (b), we showed that the upper bounds on rounding error depend inversely on h for both

$$\left|\frac{f(a+h)-f(a)}{h} - \frac{f_c(a+h)-f_c(a)}{h}\right|$$

and

$$\left|\frac{f_c(a+h)-f_c(a)}{h} - (f_c(a+h) -_{\text{algorithm}} f_c(a)) \div_{\text{algorithm}} h\right|$$

Thus, we would expect rounding error to increase as $h \to 0$, which would help explain why the rounding error dominated the absolute error for small values of h.
Also, in lecture, we showed that the upper bound on the discretization error

$$\left| f'(a) - \frac{f(a+h) - f(a)}{h} \right|$$

is directly proportional to h. Thus, we would expect rounding error to increase as $|h|$ increases. Therefore, our analysis from lecture, part (a), and part (b) combine to imply that the discretization error should be much larger than the rounding error for large h while the rounding error should be much larger than the discretization error for small h. This provides a qualitative explanation of the reason discretization error dominates for $h >> 10^{-8}$ while rounding error dominates for $h << 10^{-8}$. We can use part (d) to provide a quantitative explanation for the behavior. Since MATLAB uses IEEE standards, we know, during the creation of our graph, we have

$$\eta = \frac{1}{2}(2)^{-52} \approx 1.1 \cdot 10^{-16}$$

so, based on our calculations in part (d), the absolute error should be minimized when

$$h = 2\sqrt{\eta} = 2\sqrt{2^{-52}} \approx 2.11 \cdot 10^{-8}$$

Comparing this expected value to the graph, we see that the absolute error indeed reaches a minimum around $2 \cdot 10^{-8}$. Thus, the term that dominates the absolute error demonstrated in the graph agrees with both qualitative and quantitative expectations based on parts our analysis from parts (a), (b), and (d).
The absolute error in the graph levels off when h is very small, seemingly around $h = 10^{-16}$. This is the first point in our array h for which we have $h < \eta$, and since it is reasonable to assume $\varepsilon \approx \eta$, we have rounding error so high that $a + h = 1.2 + 10^{-16}$ evaluates to $1.2 = a$. This results in catastrophic cancellation, as MATLAB evaluates $sin(a + h) = sin(1.2 + 10^{-16})$ as equal to $sin(a) = sin(1.2)$, so it evaluates

$$|f'(a) - (f_c(a + h) -_{\text{algorithm}} f_c(a)) \div_{\text{algorithm}} h|$$

as equal to

$$|f'(a) - 0| = |f'(a)|$$

leading to constant absolute error $|f'(a)|$. As previously discussed, the rounding error already dominates the absolute error by the time h is as small as 10^{-16}, so decreasing h further cannot decrease the absolute error, despite minimizing discretization error. Thus, the overall behavior of the absolute error demonstrated in the graph, both the term that dominates it and the reason/point at which it levels off, can be explained both quantitatively and qualitatively with the analysis from parts (a) through (d).

MATH 408: Mathematical Statistics

All assignments in this section were written by Steven M. Heilman, RTPC Assistant Professor of Mathematics, USC. Solutions to assignments 1 through 6 are provided.

Assignment 1

Exercise 1.

No work requested.

Exercise 2.

No work requested.

Exercise 3.

Two people take turns throwing darts at a board. Person A goes first, and each of their throws has a probability of 1/4 of hitting the bullseye. Person B goes next, and each of their throws has a probability of 1/3 of hitting the bullseye. Then Person A goes, and so on. With what probability will Person A hit the bullseye before Person B does?

Solution.
First, note that, by the formula for an infinite Geometric Series, we have

$$\sum_{n=0}^{\infty} a \cdot r^n = \frac{a}{1-r} \qquad (1)$$

for all real-valued $|r| < 1$.

Now, let $A =$ the event that a randomly selected one of Person A's throws hits the bullseye.
Let $B =$ the event that a randomly selected one of Person B's throw's hits the bullseye.
Then $A^c =$ the event that a randomly selected one of Person A's throws misses the bullseye, and $B^c =$ the event that a randomly selected one of Person B's throws misses the bullseye.

Let $A_i =$ the event that Person A and Person B both *miss* the bullseye on their first i throws, then Person A *hits* the bullseye on their $i+1$'th throw.
This allows us to define

$A^* =$ the event that Person A hits the bullseye first $= \bigcup_{i=0}^{\infty} A_i$

Note that, if the first bullseye hit from either player occurs on Person A's $i+1$'th throw, then it could not possibly occur on Person A's $j+1$'th throw, for all $i \neq j$. This ensures that

$$A_i \cap A_j = \emptyset$$

for all $0 \leq i \neq j$, $i, j \in \mathbb{Z}$.
Applying the axiom that, for any countable set of disjoint events $A_1, ..., A_n$,

$$\mathbb{P}(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mathbb{P}(A_i)$$

we find

$$\mathbb{P}(A^*) = \mathbb{P}(\bigcup_{i=0}^{\infty} A_i) = \sum_{i=0}^{\infty} \mathbb{P}(A_i) \qquad (2)$$

so we just need to find a formula for $\mathbb{P}(A_i)$ in terms of i. Since we are given $\mathbb{P}(A) = \frac{1}{4}$ and $\mathbb{P}(B) = \frac{1}{3}$, we know

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A) = 1 - \frac{1}{4} = \frac{3}{4}$$

and

$$\mathbb{P}(B^c) = 1 - \mathbb{P}(B) = 1 - \frac{1}{3} = \frac{2}{3}$$

Assuming all throws are made independently of each other, we find the probability that Person A misses the bullseye, followed by Person B missing the bullseye is

$$\mathbb{P}(A^c \cap B^c) = \mathbb{P}(A^c)\mathbb{P}(B^c) = \frac{3}{4}\frac{2}{3} = \frac{2}{4} = \frac{1}{2}$$

Thus, the probability of Person A and Person B both missing their first i throws is

$$(\mathbb{P}(A^c \cap B^c))^i = \big(\frac{1}{2}\big)^i$$

Once again assuming the independence of throws, this implies that

$$\mathbb{P}(A_i) = (\mathbb{P}(A^c \cap B^c))^i \cdot \mathbb{P}(A) = \big(\frac{1}{2}\big)^i \cdot \frac{1}{4}$$

for all $0 \leq i \in \mathbb{Z}$. Plugging this result into (2), we find

$$\mathbb{P}(A^*) = \sum_{i=0}^{\infty} \mathbb{P}(A_i) = \sum_{i=0}^{\infty} \frac{1}{4} \cdot \big(\frac{1}{2}\big)^i$$

Applying the geometric series formula from (1) (since $r = \frac{1}{2} < 1$), we find that the probability that Perosn A hits the bullseye before Person B does is

$$\mathbb{P}(A^*) = \frac{\frac{1}{4}}{1 - \frac{1}{2}} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{4} \cdot 2 = \frac{1}{2} = 50\%$$

Thus, assuming the independence of throws, there is a 50% chance that Person A will hit the bullseye before Person B does.

Exercise 4

Suppose you have a car with twenty tires, and the car mechanic removes all twenty tires. Suppose the mechanic now puts the tires back on randomly, so that all arrangements of the tires are equally likely. With what probability will no tire end up in its original position? Give an answer to ten decimal places of accuracy (e.g. your answer could be 0.1234567891). Can you guarantee that these ten decimal places are correct?

Solution.
Let $A =$ the event that no tire ends up in its original position.
Let $B_i =$ the event that the i'th tire ends up in its original position.
Note that $A = (\bigcup_{i=1}^{20} B_i)^c$, so $\mathbb{P}(A) = 1 - \mathbb{P}(\bigcup_{i=1}^{20} B_i)$ There are 20! possible arrangements of the tires, so the sample space, Ω, has a size of $|\Omega| = 20!$. Since all possible arrangements are equally likely, we can calculate $\mathbb{P}(\bigcup_{i=1}^{20} B_i)$ by counting $|\bigcup_{i=1}^{20} B_i|$ and applying the formula

$$\mathbb{P}(\bigcup_{i=1}^{20} B_i) = \frac{|\bigcup_{i=1}^{20} B_i|}{|\Omega|} \qquad (3)$$

. To count $|\bigcup_{i=1}^{20} B_i|$, we apply the principle of inclusion exclusion to find

$$\begin{aligned}
|\bigcup_{i=1}^{20} B_i| = & \quad \sum_{i=1}^{20} \binom{20}{i}(20-i)!(-1)^{i-1} \\
= & \quad \binom{20}{1}\cdot 19! - \binom{20}{2}\cdot 18! + \binom{20}{3}\cdot 17! - \binom{20}{4}\cdot 16! + \binom{20}{5}\cdot 15! - \binom{20}{6}\cdot 14! \\
+ & \quad \binom{20}{7}\cdot 13! - \binom{20}{8}\cdot 12! + \binom{20}{9}\cdot 11! - \binom{20}{10}\cdot 10! + \binom{20}{11}\cdot 9! - \binom{20}{12}\cdot 8! \\
+ & \quad \binom{20}{13}\cdot 7! - \binom{20}{14}\cdot 6! + \binom{20}{15}\cdot 5! - \binom{20}{16}\cdot 4! + \binom{20}{17}\cdot 3! - \binom{20}{18}\cdot 2! \\
+ & \quad \binom{20}{19}\cdot 1! - \binom{20}{20}\cdot 0! \\
\approx & \quad 1.53788738\cdot 10^{18}
\end{aligned}$$

Plugging in the unrounded value of $|\bigcup_{i=1}^{20} B_i|$ into (3), we find

$$\mathbb{P}(\bigcup_{i=1}^{20} B_i) = \frac{|\bigcup_{i=1}^{20} B_i|}{|\Omega|} = \frac{1.53788738\cdot 10^{18}}{20!} = 0.6321205588$$

This allows us to compute that

$$\mathbb{P}(A) = 1 - 0.6321205588 = 0.3678794412$$

Thus, the probability that no tire ends up in its original position, expressed to ten decimal places of accuracy, is 0.3678794412.

Assuming that all arrangements of the tires are equally likely, and assuming that the calculator used did not round any results early, I can guarantee these ten decimal places are correct.

$$\sum_{i=1}^{20} \binom{20}{i}(20-i)!(-1)^{i-1}$$

is a precise quantity for the number of arrangements in which at least one tire ends up in its original place, and there are exactly 20! possible arrangements of the 20 tires, so

$$1 - \frac{\sum_{i=1}^{20} \binom{20}{i}(20-i)!(-1)^{i-1}}{20!} \qquad (4)$$

is precisely the probability that no tire ends up in its original place. Therefore, operating under the assumption that the calculator used accurately computed the expression from (4), I can guarantee that the ten decimal places in 0.3678794412 are correct.

Exercise 5.

Suppose a test for a disease is 99.9% accurate. That is, if you have the disease, the test will be positive with 99.9% probability. And if you do not have the disease, the test will be negative with 99.9% probability. Suppose also the disease is fairly rare, so that roughly 1 in 20, 000 people have the disease. If you test positive for the disease, with what probability do you actually have the disease?

Solution.
Let P = the event that you test positive for the disease.

Let $H =$ the event that you have the disease.
Then we are given

$$\mathbb{P}(H) = \frac{1}{20000}, \quad \mathbb{P}(P|H) = 0.999, \quad \mathbb{P}(P^c|H^c) = 0.999$$

and want to compute

$$\mathbb{P}(H|P)$$

By Bayes' Theorem, we know that for any two events A and B, we have

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)} \tag{5}$$

Thus, we can rewrite $\mathbb{P}(H|P)$ as

$$\mathbb{P}(H|P) = \frac{\mathbb{P}(P|H)\mathbb{P}(H)}{\mathbb{P}(P)} \tag{6}$$

We are given $\mathbb{P}(P|H)$ and $\mathbb{P}(H)$.
By the Law of Total Probability, since $H \cup H^c = \Omega$ and $H \cap H^c = \emptyset$, we have

$$\mathbb{P}(P) = \mathbb{P}(P|H)\mathbb{P}(H) + \mathbb{P}(P|H^c)\mathbb{P}(H^c) = \mathbb{P}(P|H)\mathbb{P}(H) + (1 - \mathbb{P}(P^c|H^c))\mathbb{P}(H^c) = 0.999 \cdot \frac{1}{20000} + 0.001 \cdot \frac{19999}{20000}$$

Plugging this result into (6), we find

$$\mathbb{P}(H|P) = \frac{0.999 \cdot \frac{1}{20000}}{0.999 \cdot \frac{1}{20000} + 0.001 \cdot \frac{19999}{20000}} \approx 0.047576$$

Thus, if you test positive for the disease, there is approximately a 4.76% chance that you actually have the disease.

Exercise 6.

Suppose I tell you that the following list of 20 numbers is a random sample from a Gaussian random variable, but I don't tell the mean or standard deviation.

5.1715, 3.2925, 5.2172, 6.1302, 4.9889, 5.5347, 5.2269, 4.1966, 4.7939, 3.7127 5.3884, 3.3529, 3.4311, 3.6905, 1.5557, 5.9384, 4.8252, 3.7451, 5.8703, 2.7885

To the best of your ability, determine what the mean and standard deviation are of this random variable. (This question is a bit open-ended, so there could be more than one correct way of justifying your answer.)

Solution.
Since the sample is random, we know that the sample mean is an unbiased estimator for the mean of the Gaussian random variable. Therefore, the best way to determine the mean of the Gaussian random variable X is to calculate the sample mean, $\bar{X}$. Since the size of the sample is 20, we can compute that

$$\begin{aligned} \bar{X} = \quad & \frac{1}{20}\big(5.1715 + 3.2925 + 5.2172 + 6.1302 + 4.9889 + 5.5347 + 5.2269 + 4.1966 \\ + \quad & 4.7939 + 3.7127 + 5.3884 + 3.3529 + 3.4311 + 3.6905 + 1.5557 + 5.9384 + 4.8252 \\ + \quad & 3.7451 + 5.8703 + 2.7885\big) = 4.44256 \end{aligned}$$

Since $\bar{X}$ is an unbiased estimator for the X's mean μ, this allows us to estimate that the mean of the Gaussian random variable X is

$$\mu_X \approx 4.44256$$

Unlike the mean of X, there is no unbiased estimator for the standard deviation of X, so we must rely on a corrected sample standard deviation. Since X is a Gaussian random variable, we can use the corrected sample standard deviation

$$\hat{\sigma} = \sqrt{\frac{1}{n - 1.5}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

where X_i = the i'th number from the random sample. This ensures the error of our estimate remains low with a relatively large sample of size $n = 20$. We can compute that

$$\begin{aligned}
\hat{\sigma} = \quad & \big(\frac{1}{18.5}\big((5.1715 - 4.44256)^2 + (3.2925 - 4.44256)^2 + (5.2172 - 4.44256)^2 + (6.1302 - 4.44256)^2 \\
+ \quad & (4.9889 - 4.44256)^2 + (5.5347 - 4.44256)^2 + (5.2269 - 4.44256)^2 + (4.1966 - 4.44256)^2 \\
+ \quad & (4.7939 - 4.44256)^2 + (3.7127 - 4.44256)^2 + (5.3884 - 4.44256)^2 + (3.3529 - 4.44256)^2 \\
+ \quad & (3.4311 - 4.44256)^2 + (3.6905 - 4.44256)^2 + (1.5557 - 4.44256)^2 + (5.9384 - 4.44256)^2 \\
+ \quad & (4.8252 - 4.44256)^2 + (3.7451 - 4.44256)^2 + (5.8703 - 4.44256)^2 + (2.7885 - 4.44256)^2\big)\big)^{\frac{1}{2}} \approx 1.2253
\end{aligned}$$

Thus, since X is a Gaussian random variable and we have a random sample of size $n = 20$, the best estimate for the standard deviation of X is

$$\sigma_X \approx 1.2253$$

Exercise 7.

Suppose I tell you that the following list of 20 numbers is a random sample from a Gaussian random variable, but I don't tell you the mean or standard deviation. Also, around one or two of the numbers was corrupted by noise, computational error, tabulation error, etc., so that it is totally unrelated to the actual Gaussian random variable.

-1.2045, -1.4829, -0.3616, -0.3743, -2.7298, -1.0601, -1.3298, 0.2554, 6.1865, 1.2185, -2.7273, -0.8453, -3.4282, -3.2270, -1.0137, 2.0653, -5.5393, -0.2572, −1.4512, 1.2347

To the best of your ability, determine what the mean and standard deviation are of this random variable. Supposing you had instead a billion numbers, and 5 or 10 percent of them were corrupted samples, can you come up with some automatic way of throwing out the corrupted samples? (Once again, there could be more than one right answer here; the question is intentionally open-ended.)

Solution.
Since 6.1865 and −5.5393 both differ by more than 2.5 from their closest (least different) sample value, and all other values differ by no more than 1 from their closest sample value, we assume that 6.1865 and −5.5393 are the two corrupted samples. To achieve a more accurate estimate of the mean and standard deviation, we will discard these two samples and focus on the remaining 18, presumably un-corrupted samples. This leaves us with a sample of size $n = 18$ which we assume is a valid random sample. We can now apply the same processes as in **Exercise 6** to estimate the mean and standard deviation of the Gaussian random variable, X.
For the mean, we once again rely on the unbiased estimator $\bar{X}$. Now, since $n = 18$, we have

$$\begin{aligned}
\bar{X} = \quad & \frac{1}{18}\big(-1.2045 + -1.4829 + -0.3616 + -0.3743 + -2.7298 + -1.0601 + -1.3298 \\
+ \quad & 0.2554 + 6.1865 + 1.2185 + -2.7273 + -0.8453 + -3.4282 + -3.2270 + -1.0137 \\
+ \quad & 2.0653 + -5.5393 + -0.2572 + -1.4512 + 1.2347\big) \approx -0.92883
\end{aligned}$$

Thus, since the sample of size $n = 18$ is a assumed to be a valid random sample, we have the unbiased estimate that the mean of the Gaussian random variable X is

$$\mu_X \approx -0.92883$$

Once again, we cannot find an unbiased estimate for the standard deviation of the Gaussian random variable X, even after removing the corrupted samples, so we will rely on the corrected sample standard deviation,

$$\hat{\sigma} = \sqrt{\frac{1}{n - 1.5}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

where X_i is the i'th number from our valid random sample of size $n = 18$. We can compute that

$$\begin{aligned}
\hat{\sigma} = \; & \big(\frac{1}{16.5}\big((-1.2045 - (-0.92883))^2 + (-1.4829 - (-0.92883))^2 + (-0.3616 - (-0.92883))^2 + (-0.3743 - (-0.92883))^2 \\
+ \; & (-2.7298 - (-0.92883))^2 + (-1.0601 - (-0.92883))^2 + (-1.3298 - (-0.92883))^2 + (0.2554 - (-0.92883))^2 \\
+ \; & (1.2185 - (-0.92883))^2 + (-2.7273 - (-0.92883))^2 + (-0.8453 - (-0.92883))^2 + (-3.4282 - (-0.92883))^2 \\
+ \; & (-3.2270 - (-0.92883))^2 + (-1.0137 - (-0.92883))^2 + (2.0653 - (-0.92883))^2 + (-5.5393 - (-0.92883))^2 \\
+ \; & (-0.2572 - (-0.92883))^2 + (-1.4512 - (-0.92883))^2 + (1.2347 - (-0.92883))^2\big)\big)^{\frac{1}{2}} \approx 1.549
\end{aligned}$$

Thus, since X is a Gaussian random variable, and we assume to have a random sample of size $n = 18$, the best estimate for the standard deviation of X is

$$\sigma_X \approx 1.549$$

I can come up with some automatic way of throwing out the corrupted samples. I will use the $1.5 \cdot IQR$ rule to identify and discard outliers. This involves removing any values less than $Q1 - 1.5 \cdot IQR$ or greater than $Q3 + 1.5 \cdot IQR$. Here $Q1$ is the median of the smallest $\frac{n}{2}$ numbers in the sample, while $Q3$ is the median of the largest $\frac{n}{2}$ numbers in the sample, and $IQR = Q3 - Q1$. By applying this method to **Exercise 7**, we find $Q3 = -0.0009$, $Q1 = -2.1051$, $IQR = 2.1042$, so we need to discard all values less than -5.2614 or greater than 3.1554. In this case, the method only discards the two numbers we already identified as corrupted samples. While this method is not guaranteed to remove all corrupted samples, and it will occasionally remove legitimately random (albeit unlikely) samples, it ensures that all outlying corrupted samples are removed. This should ensure the remaining sample represents the population random variable most accurately.

Assignment 2

Mathematical Statistics 408 Steven Heilman

Please provide complete and well-written solutions to the following exercises.
Due September 7, 12PM noon PST, to be uploaded as a single PDF document to Gradescope.

Homework 2 - Emerson Kahle

Exercise 1. Let $n \geq 2$ be an integer. Let $X_1, \ldots, X_n$ be a random sample of size n (that is, $X_1, \ldots, X_n$ are i.i.d. random variables). Assume that $\mu := \mathbb{E}X_1 \in \mathbb{R}$ and $\sigma := \sqrt{\mathrm{var}(X_1)} < \infty$. Let $\overline{X}$ be the sample mean and let S be the sample standard deviation of the random sample. Show the following

- $\text{Var}(\overline{X}) = \sigma^2/n$.
- $\mathbb{E}S^2 = \sigma^2$.

Solution.
First, we will show that $Var(\overline{X}) = \frac{\sigma^2}{n}$. Note that

$$Var(\overline{X}) = Var(\frac{X_1 + \cdots + X_n}{n}) \qquad (1)$$

Since $Var(aX + b) = a^2 Var(X)$ for all random variables X and constants $a, b \in \mathbb{R}$ we can simplify (1) to find:

$$Var(\overline{X}) = \frac{1}{n^2} Var(X_1 + \cdots + X_n) \qquad (2)$$

For any independent random variables X and Y, we know $Var(X+Y) = Var(X) + Var(Y)$. We can apply this fact repeatedly to (2) to find

$$Var(\overline{X}) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) \qquad (3)$$

Since $X_1, ...X_N$ are i.i.d., and we are given $\sqrt{Var(X_1)} := \sigma => Var(X_1) = \sigma^2$, we know $Var(X_i) = \sigma^2$ for all $i \in \{1, ..., n\}$. Applying this to (3), we find

$$Var(\overline{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} \cdot n\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

This completes the proof that $Var(\overline{X}) = \frac{\sigma^2}{n}$.

Now, we will show that $\mathbb{E}[S^2] = \sigma^2$. Note that

$$\mathbb{E}[S^2] = \mathbb{E}[\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2] \qquad (4)$$

Applying the fact that $\mathbb{E}[aX] = a\mathbb{E}[X]$ and Linearity of Expectation to (4), we find

$$\mathbb{E}[S^2] = \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}[(X_i - \overline{X})^2]$$

Expanding $(X_i - \overline{X})^2$ and applying Linearity of Expectation to the result, we find

$$\mathbb{E}[S^2] = \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}[X_i^2 - 2X_i\overline{X} + \overline{X}^2] = \frac{1}{n-1} \sum_{i=1}^{n} \big(\mathbb{E}[X_i^2] - 2\mathbb{E}[X_i\overline{X}] + \mathbb{E}[\overline{X}^2]\big) \qquad (5)$$

We can solve for the three expectations inside the sum separately. We will repeatedly use the fact that, for any random variable X, we have

$$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \implies Var(X) + \mathbb{E}[X]^2 = \mathbb{E}[X^2] \qquad (6)$$

We know $Var(X_i) = \sigma^2$ for all $i \in \{1, ..., n\}$ by the previous proof, and, since all X_i are identically distributed, we know

$$\mathbb{E}[X_1] = \mu \implies \mathbb{E}[X_i] = \mu$$

for all $i \in \{1, ..., n\}$. Combining this with (6), we find

$$\mathbb{E}[X_i^2] = Var(X_i) + \mathbb{E}[X_i]^2 = \sigma^2 + \mu^2 \qquad (7)$$

for all $i \in \{1, ..., n\}$.
To find $\mathbb{E}[X_i\overline{X}]$, we have to expand the product using the definition of $\overline{X}$. Note that

$$X_i\overline{X} = X_i\frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\big(X_i^2 + \sum_{j\in\{1,...,n\}s.t.j\neq i} X_iX_j\big)$$

Now, applying $\mathbb{E}[aX] = a\mathbb{E}[X]$ and Linearity of Expectation, we find

$$\mathbb{E}[X_i\overline{X}] = \mathbb{E}[\frac{1}{n}\big(X_i^2 + \sum_{j\in\{1,...,n\}s.t.j\neq i} X_iX_j)] = \frac{1}{n}\big(\mathbb{E}[X_i^2] + \sum_{j\in\{1,...,n\}s.t.j\neq i} \mathbb{E}[X_iX_j]\big) \qquad (8)$$

Note: Since we have restricted that $j \neq i$, we know that X_i and X_j are independent random variables. This allows us to use the fact that, for any independent random variables X and Y, we have

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \qquad (9)$$

Combing the result from (7), the fact from (9), and the fact that $\mathbb{E}[X_i] = \mu$ for all $i \in \{1, ..., n\}$, to (8), we find

$$\mathbb{E}[X_i\overline{X}] = \frac{1}{n}\big(\sigma^2 + \mu^2 + \sum_{j\in\{1,...,n\}s.t.j\neq i} \mathbb{E}[X_i]\mathbb{E}[X_j]\big) = \frac{1}{n}\big(\sigma^2 + \mu^2 + \sum_{j\in\{1,...,n\}s.t.j\neq i} \mu\cdot\mu\big) \qquad (10)$$

Simplifying (10) yields

$$\mathbb{E}[X_i\overline{X}] = \frac{1}{n}\big(\sigma^2 + \mu^2 + (n-1)\mu^2\big) = \frac{\sigma^2 + n\mu^2}{n} = \frac{\sigma^2}{n} + \mu^2 \qquad (11)$$

We can find the last needed expectation, $\mathbb{E}[\overline{X}]^2$, by applying (6) once more. We already proved that $Var(\overline{X}) = \frac{\sigma^2}{n}$ and we can use Linearity of Expectation to compute that

$$\mathbb{E}[\overline{X}] = \mathbb{E}[\frac{X_1 + \cdots + X_n}{n}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\cdot n\mu = \frac{n\mu}{n} = \mu \qquad (12)$$

Now, we can apply (6) to find

$$\mathbb{E}[\overline{X}^2] = Var(\overline{X}) + \mathbb{E}[\overline{X}]^2 = \frac{\sigma^2}{n} + \mu^2 \qquad (13)$$

Plugging in our results from (7), (11), and (13) into (5), we find

$$\mathbb{E}[S^2] = \frac{1}{n-1}\sum_{i=1}^{n}\big(\sigma^2 + \mu^2 - 2(\frac{\sigma^2}{n} + \mu^2) + \frac{\sigma^2}{n} + \mu^2\big) = \frac{1}{n-1}\sum_{i=1}^{n}\big(\sigma^2 - \frac{\sigma^2}{n}\big) \qquad (14)$$

Simplifying (14) yields

$$\mathbb{E}[S^2] = \frac{1}{n-1}\cdot n\cdot\frac{(n-1)\sigma^2}{n} = \sigma^2$$

This completes the proof that $\mathbb{E}[S^2] = \sigma^2$.

Exercise 2 (Optional)**.** Let $X_1, \ldots, X_n$ be i.i.d. standard Gaussian random variables (i.e. Gaussian random variables with mean zero and variance one). Show that

$$X_1^2 + \cdots + X_n^2$$

has a chi-squared distribution with n degrees of freedom.

(Hint: Let $B(0,r) := \{(x_1,\dots,x_n) \in \mathbb{R}^n \colon x_1^2 + \cdots + x_n^2 \le r^2\}$. Using hyperspherical coordinates, write

$$\begin{aligned}\mathbb{P}(X_1^2 + \cdots + X_n^2 \le t) &= (2\pi)^{-n/2} \int_{B(0,\sqrt{t})} e^{-(x_1^2+\cdots+x_n^2)/2} dx_1 \cdots dx_n \\ &= (2\pi)^{-n/2} \int_{r=0}^{r=\sqrt{t}} \int_{\partial B(0,1)} r^{n-1} e^{-r^2/2} d\sigma dr,\end{aligned}$$

where $d\sigma$ denotes integration on the boundary of the unit ball $\partial B(0,1)$. To find the latter quantity, let $t = \infty$ to note that

$$1 = (2\pi)^{-n/2} \int_{r=0}^{r=\infty} r^{n-1} e^{-r^2/2} dr \cdot \int_{\partial B(0,1)} d\sigma,$$

so that

$$\int_{\partial B(0,1)} d\sigma = \frac{(2\pi)^{n/2}}{\int_{r=0}^{r=\infty} r^{n-1} e^{-r^2/2} dr},$$

and then change variables to obtain the Gamma function on the right side denominator.)

Exercise 3. Let X be a chi squared random variables with p degrees of freedom. Let Y be a chi squared random variable with q degrees of freedom. Assume that X and Y are independent. Show that $(X/p)/(Y/q)$ has the following density, known as **Snedecor's f-distribution** with p and q degrees of freedom

$$f_{(X/p)/(Y/q)}(t) := \frac{t^{(p/2)-1}(p/q)^{p/2}\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)} \Big(1 + t(p/q)\Big)^{-(p+q)/2}, \qquad \forall\, t > 0.$$

Solution. By definition of the chi-squared distribution, we know that, since X is a chi-squared random variable with p degrees of freedom, its PDF is

$$f_X(x) = \begin{cases} \frac{x^{\frac{p}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{p}{2}}\Gamma(\frac{p}{2})} \text{ if } x > 0 \\ 0 \text{ otherwise.} \end{cases}$$

Similarly, since Y is a chi-squared random variable with q degrees of freedom, its PDF is

$$f_Y(y) = \begin{cases} \frac{y^{\frac{q}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{q}{2}}\Gamma(\frac{q}{2})} \text{ if } y > 0 \\ 0 \text{ otherwise.} \end{cases}$$

To find $f_{\frac{X/p}{Y/q}}(t)$, we will use the fact that $f_{\frac{X/p}{Y/q}}(t) = \frac{d}{dt} F_{\frac{X/p}{Y/q}}(t)$, where $F_{\frac{X/p}{Y/q}}(t)$ is the CDF of $\frac{X/p}{Y/q}$. Applying the definition of the CDF and the fact that X and Y, as chi-squared random variables, are always non-negative, we find

$$F_{\frac{X/p}{Y/q}}(t) = \mathbb{P}(\frac{X/p}{Y/q} \le t) = \mathbb{P}(\frac{X}{Y} \le \frac{tp}{q}) = \int\int_{\{(x,y)\in\mathbb{R}^2 \,|\, \frac{x}{y} \le \frac{tp}{q}, x,y>0\}} f_X(x) f_Y(y) dx dy \tag{15}$$

Since it is difficult to clearly define bounds for the region of integration, we will apply a change of variables. Define $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ as $\phi(a,b) = (ab, a) = (x,y)$, so $\phi^{-1}(x,y) = \phi^{-1}(ab,a) = (a,b)$. This implies that $a = y$ and $x = ab = yb \iff \frac{x}{y} = b$. Thus, integrating over all points (x,y) s.t. $\frac{x}{y} \le \frac{tp}{q}$, $x, y > 0$ is equivalent to integrating over all points (a,b) s.t. $b \le \frac{tp}{q}$, $a, b > 0$. This allows us to clearly define bounds on a and b as $0 < a \le \infty$ and $0 < b \le \frac{tp}{q}$. To complete the change of variables, we recall the general change of variables formula,

$$\int\int_{\phi(U)} f(x,y) dx dy = \int\int_U f(\phi(a,b)) |Jac(\phi(a,b))| da db$$

Here, $f_X(x)f_Y(y) = f(x,y)$,

$\{(x,y) \in \mathbb{R}^2 | \frac{x}{y} \leq \frac{tp}{q}, x, y > 0\} = \phi(\{(a,b) \in \mathbb{R}^2 | a, b > 0,\, b \leq \frac{tp}{q}\})$, and

$$|Jac(\phi(a,b))| = |Jac(ab,a)| = |det(\begin{matrix} \frac{\partial(ab)}{\partial a} & \frac{\partial(ab)}{\partial b} \\ \frac{\partial(a)}{\partial a} & \frac{\partial(a)}{\partial b} \end{matrix})| = |det(\begin{matrix} b & a \\ 1 & 0 \end{matrix})| = |0 - a| = |-a| = |a|$$

Plugging these values into the change of variables formula and combining with (15), we find

$$F_{\frac{X/p}{Y/q}}(t) = \int\int_{\{(a,b) \in \mathbb{R}^2 | a,b>0,\, b \leq \frac{tp}{q}\}} f_X(ab) f_Y(a) |a| da db = \int_{b=0}^{\frac{tp}{q}} \int_{a=0}^{\infty} f_X(ab) f_Y(a) |a| da db \tag{16}$$

Since we are integrating w.r.t a from 0 to ∞, we can replace $|a|$ with a in (16) since $a = |a|$ for all $0 < a \leq \infty$. Then, we can apply the fact that $f_{\frac{X/p}{Y/q}}(t) = \frac{d}{dt} F_{\frac{X/p}{Y/q}}(t)$ and the Fundamental Theorem of Calculus to find

$$f_{\frac{X/p}{Y/q}}(t) = \frac{d}{dt} F_{\frac{X/p}{Y/q}}(t) = \frac{d}{dt} \int_{b=0}^{\frac{tp}{q}} \int_{a=0}^{\infty} a f_X(ab) f_Y(a) da db = \frac{p}{q} \int_{a=0}^{\infty} a f_X(a(\frac{tp}{q})) f_Y(a) da \tag{17}$$

with the constant $\frac{p}{q}$ appearing as the result of $\frac{d}{dt}(\frac{tp}{q})$.
Now, we can plug in the previously established PDFs of X and Y to find

$$f_{\frac{X/p}{Y/q}}(t) = \frac{p}{q} \int_{a=0}^{\infty} a \cdot \frac{(\frac{atp}{q})^{\frac{p}{2}-1} e^{-\frac{atp}{2q}}}{2^{\frac{p}{2}} \Gamma(\frac{p}{2})} \cdot \frac{a^{\frac{q}{2}-1} e^{-\frac{a}{2}}}{2^{\frac{q}{2}} \Gamma(\frac{q}{2})} da \tag{18}$$

Pulling out constants from (18) and simplifying yields

$$f_{\frac{X/p}{Y/q}}(t) = \frac{\frac{p}{q} \cdot \frac{tp}{q}^{\frac{p}{2}-1}}{2^{\frac{p+q}{2}} \cdot \Gamma(\frac{p}{2}) \cdot \Gamma(\frac{q}{2})} \int_{a=0}^{\infty} a \cdot a^{\frac{p}{2}-1} \cdot a^{\frac{q}{2}-1} \cdot e^{-\frac{atp}{2q}} \cdot e^{-\frac{a}{2}} da = \frac{(\frac{p}{q})^{\frac{p}{2}} t^{\frac{p}{2}-1}}{2^{\frac{p+q}{2}} \Gamma(\frac{p}{2}) \Gamma(\frac{q}{2})} \int_{a=0}^{\infty} a^{\frac{p+q}{2}-1} e^{-\frac{a(1+\frac{tp}{q})}{2}} da \tag{19}$$

Note: By definition, a Gamma random variable G with parameters α and β has PDF

$$f_G(x) := \begin{cases} \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha} \Gamma(\alpha)} \text{ if } x > 0 \\ 0 \text{ otherwise.} \end{cases}$$

and

$$\mathbb{P}(0 < G < \infty) = \int_0^{\infty} f_G(x) dx = 1$$

If we let $\alpha = \frac{p+q}{2}$ and $\beta = \frac{2}{(1+\frac{tp}{q})}$, we see the PDF becomes

$$f_G(x) := \begin{cases} \frac{x^{\frac{p+q}{2}-1} e^{-\frac{x(1+\frac{tp}{q})}{2}}}{\beta^{\alpha} \Gamma(\alpha)} \text{ if } x > 0 \\ 0 \text{ otherwise.} \end{cases}$$

Comparing this to (19), we see that the integrand is simply the density of a Gamma random variable multiplied by $\beta^{\alpha}\Gamma(\alpha)$. That is

$$a^{\frac{p+q}{2}-1} e^{-\frac{a(1+\frac{tp}{q})}{2}} = \beta^{\alpha} \Gamma(\alpha) f_G(a)$$

where $\alpha = \frac{p+q}{2}$ and $\beta = \frac{2}{(1+\frac{tp}{q})}$.
Plugging this result into (19), we find

$$f_{\frac{X/p}{Y/q}}(t) = \frac{t^{\frac{p}{2}-1} (\frac{p}{q})^{\frac{p}{2}}}{2^{\frac{p+q}{2}} \Gamma(\frac{p}{2}) \Gamma(\frac{q}{2})} \cdot \beta^{\alpha} \cdot \Gamma(\alpha) \cdot \int_0^{\infty} f_G(a) da = \frac{t^{\frac{p}{2}-1} (\frac{p}{q})^{\frac{p}{2}}}{2^{\frac{p+q}{2}} \Gamma(\frac{p}{2}) \Gamma(\frac{q}{2})} \cdot (\frac{2}{(1+\frac{tp}{q})})^{\frac{p+q}{2}} \cdot \Gamma(\frac{p+q}{2}) \tag{20}$$

Simplifying (20) yields

$$f_{\frac{X/p}{Y/q}}(t) = \frac{t^{\frac{p}{2}-1}(\frac{p}{q})^{\frac{p}{2}}\Gamma(\frac{p+q}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})}(1+\frac{tp}{q})^{-\frac{p+q}{2}}$$

for all $t > 0$. This completes the proof that

$$f_{(X/p)/(Y/q)}(t) := \frac{t^{(p/2)-1}(p/q)^{p/2}\Gamma((p+q)/2)}{\Gamma(p/2)\Gamma(q/2)}\Big(1+t(p/q)\Big)^{-(p+q)/2}, \qquad \forall\, t > 0.$$

Exercise 4 (Order Statistics)**.** Let $X\colon \Omega \to \mathbb{R}$ be a random variable. Let $X_1, \ldots, X_n$ be a random sample of size n from X. Define $X_{(1)} := \min_{1\le i\le n} X_i$, and for any $2 \le i \le n$, inductively define

$$X_{(i)} := \min\Big\{\{X_1, \ldots, X_n\} \smallsetminus \{X_{(1)}, \ldots, X_{(i-1)}\}\Big\},$$

so that

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)} = \max_{1\le i\le n} X_i.$$

The random variables $X_{(1)}, \ldots, X_{(n)}$ are called the **order statistics** of $X_1, \ldots, X_n$.

- Suppose X is a discrete random variable and we can order the values that X takes as $x_1 < x_2 < \cdots$. For any $i \ge 1$, define $p_i := \mathbb{P}(X \le x_i)$. Show that, for any $1 \le i, j \le n$,

$$\mathbb{P}(X_{(j)} \le x_i) = \sum_{k=j}^{n} \binom{n}{k} p_i^k (1-p_i)^{n-k}.$$

(Hint: Let Y be the number of indices $1 \le j \le n$ such that $X_j \le x_i$. Then Y is a binomial random variable with parameters n and p_i.)

You don't have to show it, but if X is a continuous random variable with density f_X and cumulative distribution function F_X, then for any $1 \le j \le n$, $F_{X_{(j)}}$ has density

$$f_{X_{(j)}}(x) := \frac{n!}{(j-1)!(n-j)!} f_X(x)(F_X(x))^{j-1}(1-F_X(x))^{n-j}, \qquad \forall\, x \in \mathbb{R}.$$

(This follows by differentiating the above identity for the cumulative distribution function, i.e. by differentiating $\mathbb{P}(X_{(j)} \le x) = \sum_{k=j}^{n} \binom{n}{k} F_X(x)^k (1-F_X(x))^{n-k}$, where $F_X(x) := \mathbb{P}(X \le x)$ for any $x \in \mathbb{R}$.)

Solution.
Let $Y =$ the number of indices $1 \le j \le n$ s.t. $X_j \le x_i$. Then $Y \sim \mathit{Binomial}(n, p_i)$. Note that $X_{(j)}$ is the jth smallest item from our sample of size n. Thus, $X_{(j)} \le x_i$ is only possible if $\exists \ge j$ items from the sample that are $\le x_i$. Note that Y is exactly the number of items from the sample that are smaller than x_i. Moreover, if $\exists \ge j$ items from the sample that are $\le x_i$ (i.e. if $Y \ge j$), then since $X_{(j)}$ is the jth smallest item, we are guaranteed to have $X_{(j)} \le X_{(Y)} \le x_i$. Therefore, $(X_j \le x_i) \iff (Y \ge j)$. This allows us to rewrite the probability in question as

$$\mathbb{P}(X_{(j)} \le x_i) = \mathbb{P}(Y \ge j) \qquad (21)$$

Since $Y \sim \mathit{Binomial}(n, p_i)$, we know it's CDF is

$$F_Y(x) = \mathbb{P}(Y \le x) = \sum_{k=0}^{x} \binom{n}{k} p_i^k (1-p_i)^{n-k}$$

and we can easily compute that

$$\begin{aligned}
\mathbb{P}(Y \geq x) = &\quad 1 - F_Y(x) + \mathbb{P}(Y = x) = 1^n - F_Y(x) + \mathbb{P}(Y = x) \\
= &\quad (p_i + (1 - p_i))^n - F_Y(x) + \binom{n}{x} p_i^x (1 - p_i)^{n-x} \\
=_* &\quad \sum_{k=0}^{n} \binom{n}{k} p_i^k (1 - p_i)^{n-k} - \sum_{k=0}^{x} \binom{n}{k} p_i^k (1 - p_i)^{n-k} + \binom{n}{x} p_i^x (1 - p_i)^{n-x} \\
= &\quad \sum_{k=x}^{n} \binom{n}{k} p_i^k (1 - p_i)^{n-k} \qquad (22)
\end{aligned}$$

With the $=_*$ indicating that the equality follows from the Binomial Theorem. Plugging the result from (22) into (21) with x = j, we find

$$\mathbb{P}(X_{(j)} \leq x_i) = \mathbb{P}(Y \geq j) = \sum_{k=j}^{n} \binom{n}{k} p_i^k (1 - p_i)^{n-k}$$

This completes the proof that, for any $1 \leq i, j \leq n$,

$$\mathbb{P}(X_{(j)} \leq x_i) = \sum_{k=j}^{n} \binom{n}{k} p_i^k (1 - p_i)^{n-k}.$$

- Let X be a random variable uniformly distributed in $[0, 1]$. For any $1 \leq j \leq n$, show that $X_{(j)}$ is a beta distributed random variable with parameters j and $n - j + 1$. Conclude that (as you might anticipate)

$$\mathbb{E}X_{(j)} = \frac{j}{n+1}.$$

Solution.
Since $X \sim ContinuousUniform([0, 1])$, we are given that

$$f_{X_{(j)}}(x) := \frac{n!}{(j-1)!(n-j)!} f_X(x)(F_X(x))^{j-1}(1 - F_X(x))^{n-j}, \qquad \forall\, x \in \mathbb{R}.$$

To classify $X_{(j)}$ as a Beta distributed random variable with parameters j and $n - j + 1$, we first need to find $F_X(x)$ and $f_X(x)$. Since $X \sim ContinuousUniform([0, 1])$, we know that

$$F_X(x) = \mathbb{P}(X \leq x) = \begin{cases} 1 \text{ if } x \geq 1 \\ x \text{ if } 0 < x < 1 \\ 0 \text{ otherwise.} \end{cases} \qquad (23)$$

Applying the fact that $\frac{d}{dt} F_X(x) = f_X(x)$ to (23) yields

$$f_X(x) = \begin{cases} 1 \text{ if } 0 < x < 1 \\ 0 \text{ otherwise} \end{cases} \qquad (24)$$

Plugging the results from (23) and (24) into the definition of $f_{X_{(j)}}$ yields

$$f_{X_{(j)}}(x) = \begin{cases} \frac{n!}{(j-1)!(n-j)!} x^{j-1}(1 - x)^{n-j} \text{ if } 0 < x < 1 \\ 0 \text{ otherwise} \end{cases}$$

Claim: For all $n \geq 1$, $n \in \mathbb{Z}$, we have

$$\Gamma(n) = (n-1)!$$

Proof: We apply mathematical induction on n.
Base Case: $n = 1$, we have

$$\Gamma(1) = \int_0^\infty x^0 e^{-x} dx = -e^{-x}|_0^\infty = 0 + 1 = 1 = 0!;$$

Inductive Hypothesis: Assume $\Gamma(n) = (n-1)!$ for all $1 \leq n \leq k$.
Inductive Step: Consider $\Gamma(k+1)$. Let $dv = e^{-x}dx$, $v = -e^{-x}$, $u = x^k$, $du = kx^{k-1}$. Then integrate by parts to find

$$\Gamma(k+1) = \int_0^\infty x^k e^{-x} dx = -x^k e^{-x}|_0^\infty + k\int_0^\infty x^{k-1} e^{-x} dx = 0 + k\Gamma(k) = k\Gamma(k) \tag{25}$$

By the Inductive Hypothesis, we know $\Gamma(k) = (k-1)!$. Plugging this result into (25), we find

$$\Gamma(k+1) = k(k-1)! = k!$$

The conclusion that $\Gamma(n) = (n-1)!$ follows for all $1 \leq n \in \mathbb{Z}$ by induction.

We can now rewrite $f_{X_{(j)}}$ as

$$f_{X_{(j)}}(x) = \begin{cases} \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} x^{j-1}(1-x)^{n-j} \text{ if } 0 < x < 1 \\ 0 \text{ otherwise.} \end{cases} \tag{26}$$

By definition, the PDF of the Beta distribution is

$$f(x) := \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \text{ if } 0 < x < 1 \\ 0 \text{ otherwise.} \end{cases}$$

where $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}$.
Note: From lecture notes, we know that

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

If we let $\alpha = j$ and $\beta = n - j + 1$, this identity becomes

$$B(j, n-j+1) = \frac{\Gamma(j)\Gamma(n-j+1)}{\Gamma(n+1)}$$

Applying this identity to (26), we find that $X_{(j)}$ has density function

$$f_{X_{(j)}}(x) = \begin{cases} \frac{1}{B(j,n-j+1)} x^{j-1}(1-x)^{(n-j+1)-1} \text{ if } 0 < x < 1 \\ 0 \text{ otherwise.} \end{cases} \tag{27}$$

which is precisely the PDF of a Beta distributed random variable with parameters $\alpha = j$, $\beta = n-j+1$. This concludes the proof that $X_{(j)}$ is a Beta distributed random variable with parameters $\alpha = j$ and $\beta = n - j + 1$.

We can directly compute that, by definition,

$$\mathbb{E}[X_{(j)}] = \int_{-\infty}^\infty x f_{X_{(j)}}(x) dx = \frac{1}{B(j, n-j+1)} \int_0^1 x^j (1-x)^{n-j} = \frac{B(j+1, n-j+1)}{B(j, n-j+1)} \tag{28}$$

since $\int_0^1 x^j(1-x)^{n-j} := B(j+1, n-j+1)$. Applying the identity that $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ to (28) yields

$$\mathbb{E}[X_{(j)}] = \frac{\Gamma(j+1)\Gamma(n-j+1)\Gamma(n+1)}{\Gamma(n+2)\Gamma(j)\Gamma(n-j+1)} = \frac{\Gamma(j+1)\Gamma(n+1)}{\Gamma(n+2)\Gamma(j)} \tag{29}$$

Applying $\Gamma(n) = (n-1)!$ for all $1 \leq n \in \mathbb{Z}$, we find

$$\mathbb{E}[X_{(j)}] = \frac{j!n!}{(n+1)!(j-1)!} = \frac{j}{n+1}$$

This completes the proof that $\mathbb{E}[X_{(j)}] = \frac{j}{n+1}$.

- Let $a, b \in \mathbb{R}$ with $a < b$. Let U be the number of indices $1 \leq j \leq n$ such that $X_j \leq a$. Let V be the number of indices $1 \leq j \leq n$ such that $a < X_j \leq b$. Show that the vector $(U, V, n-U-V)$ is a multinomial random variable, so that for any nonnegative integers u, v with $u + v \leq n$, we have

$$\begin{aligned}&\mathbb{P}(U = u, V = v, n - U - V = n - u - v)\\ &\quad = \frac{n!}{u!v!(n-u-v)!} F_X(a)^u (F_X(b) - F_X(a))^v (1 - F_X(b))^{n-u-v}.\end{aligned}$$

Consequently, for any $1 \leq i, j \leq n$,

$$\mathbb{P}(X_{(i)} \leq a, X_{(j)} \leq b) = \mathbb{P}(U \geq i, U + V \geq j) = \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} \mathbb{P}(U = k, V = m) + \mathbb{P}(U \geq j).$$

So, it is possible to write an explicit formula for the joint distribution of $X_{(i)}$ and $X_{(j)}$ (but you don't have to write it yourself).

Solution.
By definition, a multinomial random variable describes a situation of n independent trials, each of which falls into one of k mutually disjoint categories with a fixed probability p_i for the ith category for all $1 \leq i \leq k, i \in \mathbb{Z}$. In this case, our n independent trials are $\{X_1, ..., X_n\}$, $k = 3$, and our three categories into exactly one of which each X_i must fall are

$$\begin{cases} X_i \leq a \\ a < X_i \leq b \\ X_i > b \end{cases}$$

The fixed probabilities that correspond to each of these categories are

$$\begin{cases} \mathbb{P}(X_i \leq a) = F_X(a) \\ \mathbb{P}(a < X_i \leq b) = F_X(b) - F_X(a) \\ \mathbb{P}(X_i > b) = 1 - F_X(b) \end{cases}$$

Note that U counts the number of items from the sample of size n that fall into the $X_i \leq a$ category, V counts the number that fall into the $a < X_i \leq b$ category, and $n - U - V$, as the number of remaining items from the sample, counts the number that fall into the $X_i > b$ category. Thus, the vector $(U, V, n-U-V)$ perfectly matches the description of a multinomial random variable with 3 categories and fixed category probabilities $F_X(a)$, $F_X(b) - F_X(a)$, and $1 - F_X(b)$ respectively. Since the PMF of a multinomial random variable is

$$\mathbb{P}(N_1 = n_1, ..., N_k = n_k) = \binom{n}{n_1, ..., n_k} \prod_{i=1}^{k} p_i^{n_i}$$

where $p_i = \mathbb{P}(X_j \text{ is in the } i\text{th category})$ for some randomly selected X_j, we know

$$\begin{aligned}\mathbb{P}(U=u, V=v, n-U-V=n-u-v) &= \binom{n}{u,v,n-u-v} F_x(a)^u (F_X(b)-F_X(a))^v (1-F_X(b))^{n-u-v} \\ &= \frac{n!}{u!v!(n-u-v)!} F_x(a)^u (F_X(b)-F_X(a))^v (1-F_X(b))^{n-u-v}\end{aligned}$$

We can also directly compute $\mathbb{P}(U=u, V=v, n-U-V=n-u-v)$ to conclude that $(U, V, n-U-V)$ is a multinomial random variable. Note that, for $(U=u, V=v, n-U-V=n-u-v)$ to be true under the restrictions $u, v \geq 0, u+v \leq n$, we need to have three sets A, B, and C s.t. $A \cup B \cup C = \{X_1, ..., X_n\}$, $|A| = u$, $|B| = v$, $|C| = n-u-v$, $\forall X_i \in A, X_i \leq a$, $\forall X_i \in B, a < X_i \leq b$, and for all $X_i \in C, b < X_i$. There are $\binom{n}{u,v,n-u-v}$ ways to split $\{X_1, ..., X_n\}$ into sets A, B and C of size u, v, and $n-u-v$ respectively. Since these groupings are all equally likely, it suffices to find the probability that an arbitrary such grouping will result in $U=u, V=v, n-U-V=n-u-v$, then multiply it by the number of possible such groupings.

For any such grouping of $\{X_1, ..., X_n\}$, we need all $X_i \in A$ to satisfy $X_i \leq a$, all $X_i \in B$ to satisfy $a < X_i \leq b$, and all $X_i \in C$ to satisfy $b < X_i$. Since all X_i are sampled independently of each other, we have

$$\mathbb{P}(X_i \leq a) = F_X(a)$$

for all $X_i \in A$. Thus,

$$\mathbb{P}(X_i \leq a \forall X \in A) = \prod_{i=1}^{u} F_X(a) = F_X(a)^u$$

Similarly, we have

$$\mathbb{P}(a < X_i \leq b) = F_X(b) - F_X(a)$$

for all $X_i \in B$ which implies

$$\mathbb{P}(a < X_i \leq b \forall X_i \in B) = \prod_{i=1}^{v} (F_X(b) - F_X(a)) = (F_X(b) - F_X(a))^v$$

and

$$\mathbb{P}(X_i > b) = 1 - F_X(b)$$

for all $X_i \in C$, which implies

$$\mathbb{P}(X_i > b \forall X_i \in C) = \prod_{i=1}^{n-u-v} (1 - F_X(b))^{n-u-v}$$

Once again, since all X_i are sampled randomly and independently, we know

$$\mathbb{P}(X_i \leq a \forall X_i \in A, a < X_i \leq b \forall X_i \in B, X_i > b \forall X_i \in C) = F_X(a)^u (F_X(b) - F_X(a))^v (1 - F_X(b))^{n-u-v}$$

for any grouping of $\{X_1, ..., X_n\}$ into sets A, B, and C of size u, v, and $n-u-v$ respectively. Since there are $\binom{n}{u,v,n-u-v}$ such groupings, we know

$$\mathbb{P}(U=u, V=v, n-U-V=n-u-v) = \binom{n}{u,v,n-u-v} F_X(a)^u (F_X(b)-F_X(a))^v (1-F_X(b))^{n-u-v}$$

Applying the fact that $\binom{n}{n_1,...,n_k} = \frac{n!}{n_1!\cdots n_k!}$, we find

$$\mathbb{P}(U=u, V=v, n-U-V=n-u-v) = \frac{n!}{u!v!(n-u-v)!} F_X(a)^u (F_X(b)-F_X(a))^v (1-F_X(b))^{n-u-v}$$

This is the PMF of a multinomial random variable, which completes the computational proof that $(U, V, n-U-V)$ is a multinomial random variable with PMF

$$\mathbb{P}(U=u, V=v, n-U-V=n-u-v) = \frac{n!}{u!v!(n-u-v)!} F_X(a)^u (F_X(b)-F_X(a))^v (1-F_X(b))^{n-u-v}$$

Now, we will explain the consequence of this result.
Note that $(X_{(i)} \leq a, X_{(j)} \leq b)$ is only possible if there are *at least* i items in the sample $\leq a$ and *at least* j items in the sample $\leq b$. Also note that U counts the number of items in the sample $\leq a$ and V counts the number of items in the sample between a and b, so $U+V$ counts the number of items in the sample $\leq b$. Moreover, if $U \geq i$ and $U+V \geq j$, we are guaranteed to have $X_{(i)} \leq X_U \leq a$ and $X_{(j)} \leq X_{U+V} \leq b$. Thus, we know

$$(X_{(i)} \leq a, X_{(j)} \leq b) \iff (U \geq i, U+V \geq j)$$

so we can write

$$\mathbb{P}(X_{(i)} \leq a, X_{(j)} \leq b) = \mathbb{P}(U \geq i, U+V \geq j)$$

For $(U \geq i, U+V \geq j)$ to be true, either $U \geq j$ or $i \leq U \leq j-1$ and $j-U \leq V \leq n-U$. These are mutually disjoint events, so we can sum their probabilities to compute $\mathbb{P}(U \geq i, U+V \geq j)$. We can compute $\mathbb{P}(U \geq j)$ directly, and we can sum over all $\{(k,m) | i \leq k \leq j-1, j-k \leq m \leq n-k\}$ to find

$$\mathbb{P}(U \geq i, U+V \geq j) = \sum_{k=i}^{j-1} \sum_{m=j-k}^{n-k} \mathbb{P}(U=k, V=m) + \mathbb{P}(U \geq j)$$

This completes the explanation of the consequence.

Remark 1. We might occasionally do some computer-based exercises. You can use whatever program you want to do these exercises. Here are some links for downloading such software:

Matlab software download
R software download

Exercise 5. Using Matlab (or any other mathematical system on a computer), verify that its random number generator agrees with the law of large numbers and central limit theorem. For example, average 10^7 samples from the uniform distribution on $[0,1]$ and check how close the sample average is to $1/2$. Then, make a histogram of 10^7 samples from the uniform distribution on $[0,1]$ and check how close the histogram is to a Gaussian.

Solution.
First, we will verify that the random number generator agrees with the Law of Large Numbers. The (Weak) Law of Large Numbers states that, for any *i.i.d.* random variables $X_1, ..., X_n$ with finite mean $\mu := \mathbb{E}[X_i] < \infty$, and any $\varepsilon > 0$, we have

$$\lim_{n \to \infty} \mathbb{P}\left(\left| \frac{X_1 + \cdots + X_n}{n} - \mu \right| > \varepsilon \right) = 0$$

We sample from the Uniform([0,1]) distribution, which has an expected value of

$$\mu := \mathbb{E}[X] = \frac{0+1}{2} = \frac{1}{2}$$

Thus, the law of large numbers tells us that, since our sample of size 10^7 is so large, the chance that there is a significant difference between our sample mean and the true mean $\mu = \frac{1}{2}$ approaches 0. Our simulation of a size 10^7 random sample from Uniform([0,1]) on Matlab agrees with this law, as the sample average from our sample of size 10^7 was exactly $\overline{X}_{10^7} = 0.5000 = \frac{1}{2}$. Therefore, Matlab's random number generator agrees with the Weak Law of Large Numbers.

Now, we will show that the random number generator agrees with the Central Limit Theorem. The Central Limit Theorem tells us that, for *i.i.d.* random variables $X_1, ..., X_n$ with mean $\mu = \mathbb{E}[X_i] < \infty$ and variance $0 < Var(X_i) = \sigma^2 < \infty$, then for any $-\infty \leq a \leq \infty$, we have

$$\lim_{n\to\infty} \mathbb{P}(\frac{X_1 + \cdots X_n - n\mu}{\sigma\sqrt{n}} \leq a) = \int_{-\infty}^{a} e^{-\frac{t^2}{2}} \frac{dt}{\sqrt{2\pi}}$$

which implies that $\frac{X_1+\cdots X_n-n\mu}{\sigma\sqrt{n}}$ converges in distribution to a $Gaussian(0,1)$ random variable. Note that

$$\overline{X} = \frac{(\sigma\sqrt{n}(\frac{X_1+\cdots X_n-n\mu}{\sigma\sqrt{n}}) + n\mu)}{n}$$

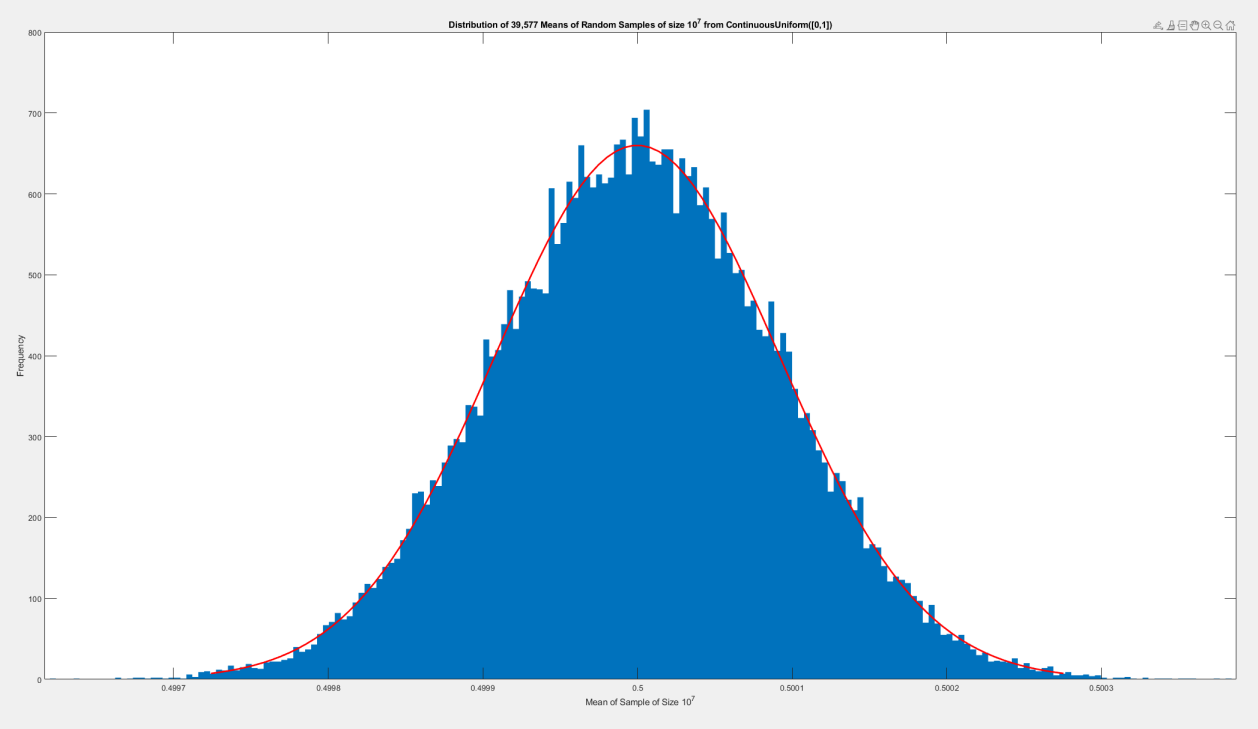
so the CLT tells us that $\overline{X}$ also converges in distribution to a Gaussian random variable as $n \to \infty$ with

$$\mathbb{E}[\overline{X}] = \frac{1}{n} \cdot (0 + n\mu) = \frac{n\mu}{n} = \mu$$

and

$$Var(\overline{X}) = \frac{1}{n^2} \cdot \sigma^2 n \cdot 1 = \frac{\sigma^2 n}{n^2} = \frac{\sigma^2}{n}$$

Thus, to test whether the Matlab random number generator agrees with the Central Limit Theorem, we take 39,577 random samples, each of size $n = 10^7$, and plot a histogram of the 39,577 corresponding $\overline{X}_n$'s. We also plot a Gaussian distribution over the histogram in orange to better judge the fit of the sample data. The results of the histogram are displayed here:



As the histogram shows, the sample means follow the Gaussian distribution very closely, as the CLT leads us to expect. Thus, Matlab's random number generator agrees with both the Weak Law of Large Numbers and the Central Limit Theorem.

Exercise 6 (Sunspot Data)**.** This exercise deals with sunspot data from the following files (the same data appears in different formats)

txt file csv (excel) file
These files are taken from http://www.sidc.be/silso/datafiles#total
To work with this data, e.g. in Matlab you can use the command

```
x=importdata('SN_d_tot_V2.0.txt')
```

to import the .txt file.

The format of the data is as follows.

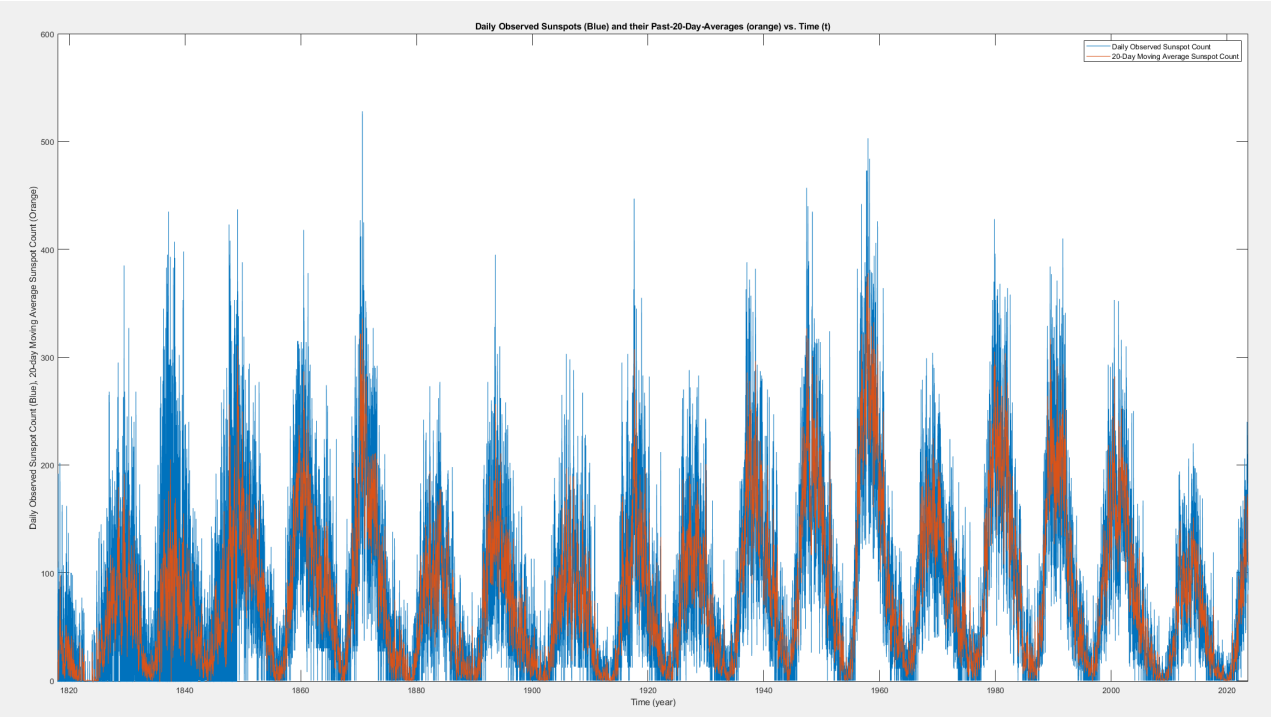
- Columns 1-3: Gregorian calendar date (Year, Month, then Day)
- Column 4: Date in fraction of year
- Column 5: Daily total number of sunspots observed on the sun. A value of -1 indicates that no number is available for that day (missing value).
- Column 6: Daily standard deviation of the input sunspot numbers from individual stations.
- Column 7: Number of observations used to compute the daily value.
- Column 8: Definitive/provisional indicator. A blank indicates that the value is definitive. A '*' symbol indicates that the value is still provisional and is subject to a possible revision (Usually the last 3 to 6 months)

For this data set, do the following:

• Plot the number of sunspots U_t versus time t. Label and scale the axes appropriately. On this same plot, also plot some moving averages of U_t. For example, for a given time t, plot the average of the twenty previous days' sunspot counts, versus time t.

Solution.

Below are the results of the time plot of Daily Observed Sunspots U_t and 20-day Moving Averages of Daily Observed Sunspots versus time t. The Daily Observed counts appear in blue, while the 20-day Moving Averages appear in orange.



- Find the sample average and sample standard deviation of U_t, averaging over all t given in the data.

Solution. We calculated the sample average and sample standard deviation using the mean(A) and std(A) functions in Matlab, where A was the entire array/vector of daily observed sunspots. The results were a sample mean of $\overline{X}_{75118} = 78.7893$ and a sample standard deviation of $S = 77.1698$, with these values maintaining accuracy up to the rounding error of the Matlab software.

- Do you notice any periodic behavior in U_t versus t?

Solution.
I do notice periodic behavior in U_t versus t. It appears as though, approximately every 10-12 years, the Daily Sunspot Observations peak at an average around 300, then dip to an average less than 50, then rise back to an average around 300, at which point the cycle repeats. This cycle appears consistent, albeit with alterations to the specific averages at different points in time, over the entire recorded time range from 1818 to the 2023. The existence and consistency of this cyclical pattern suggests a relatively strong relationship between time t and Daily Sunspot Observations U_t.

Assignment 3

Mathematical Statistics 408 Steven Heilman

Please provide complete and well-written solutions to the following exercises.
Due September 21, 12PM noon PST, to be uploaded as a single PDF document to Gradescope.

Homework 3 - Emerson Kahle

Exercise 7. Recall that a gamma distributed random variable X with parameters $\alpha, \beta > 0$ satisfies

$$\mathbb{E}X = \alpha\beta, \qquad \mathbb{E}X^2 = \alpha\beta^2.$$

Find the method of moments estimator for α, and also find the method of moments estimator for β.

Solution.
Since we have two unknowns, α and β, and we have two equations in terms of moments of X and our unknowns α and β, we can solve the system of equations to express α and β in terms of the first and second moment of X:

$$\begin{aligned}
\mathbb{E}[X] = &\qquad \alpha\beta \qquad (1)\\
\mathbb{E}[X^2] = &\qquad \alpha\beta^2 \qquad (2)\\
(1) \implies &\qquad \frac{\mathbb{E}[X]}{\beta} = \alpha \qquad (3)\\
(2),(3) \implies &\qquad \mathbb{E}[X^2] = \frac{\mathbb{E}[X]}{\beta}\beta^2 = \mathbb{E}[X]\beta \qquad (4)\\
(4) \implies &\qquad \beta = \frac{\mathbb{E}[X^2]}{\mathbb{E}[X]} \qquad (5)\\
(3),(5) \implies &\qquad \alpha = \frac{\mathbb{E}[X]}{\frac{\mathbb{E}[X^2]}{\mathbb{E}[X]}} = \frac{\mathbb{E}[X]^2}{\mathbb{E}[X^2]} \qquad (6)
\end{aligned}$$

Now that we have expressed α and β in terms of $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$, we can substitute the sample moments for $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ into (5) and (6) to find the method of moments estimators for α and β.
First, define $X_1, \cdots, X_n$ to be *i.i.d.* gamma distributed random variables with parameters $\alpha, \beta > 0$. Then the first sample moment of X is

$$M_1(X_1, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and the second sample moment is

$$M_2(X_1, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Substituting $M_1(X_1, ..., X_n)$ for $\mathbb{E}[X]$ and $M_2(X_1, ..., X_n)$ for $\mathbb{E}[X^2]$ into (5), we can define our estimator for β to be

$$\mathcal{B} := \frac{M_2(X_1, ..., X_n)}{M_1(X_1, ..., X_n)} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{\frac{1}{n}\sum_{i=1}^{n} X_i} = \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i} \qquad (7)$$

Making the same substitutions into (6) similarly allows us to define our estimator for α to be

$$\mathcal{A} = \frac{M_1(X_1, ..., X_n)}{\mathcal{B}} = \frac{(\frac{1}{n}\sum_{i=1}^{n} X_i)^2}{\frac{1}{n}\sum_{i=1}^{n} X_i^2} \qquad (8)$$

Thus, the method of moments estimators for α and β are

$$\mathcal{A} := \frac{(\frac{1}{n}\sum_{i=1}^{n} X_i)^2}{\frac{1}{n}\sum_{i=1}^{n} X_i^2}$$

and

$$\mathcal{B} := \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i}$$

respectively.

Exercise 8. Let $\sigma > 0$, and suppose a random variable X has density

$$f(x) := \frac{1}{2\sigma} e^{-|x|/\sigma}, \qquad \forall\, x \in \mathbb{R}.$$

Find the method of moments estimator for σ.

Solution.
First, we need to express σ in terms of the moments of X. For the first moment of X, we find that

$$\mathbb{E}[X] = \int_{\infty}^{\infty} x f(x) dx = \int_{\infty}^{\infty} \frac{x}{2\sigma} e^{\frac{-|x|}{\sigma}} dx \qquad (9)$$

Note that, for all $x \in \mathbb{R}$,

$$\frac{-x}{2\sigma} e^{\frac{-|-x|}{\sigma}} = \frac{-x}{2\sigma} e^{\frac{-|x|}{\sigma}} = -\frac{x}{2\sigma} e^{\frac{-|x|}{\sigma}}$$

so $\frac{x}{2\sigma} e^{\frac{-|x|}{\sigma}}$ is an *odd* function, so we know that

$$\int_0^{\infty} \frac{x}{2\sigma} e^{\frac{-|x|}{\sigma}} dx = -\int_0^{\infty} \frac{-x}{2\sigma} e^{\frac{-|-x|}{\sigma}} dx = -\int_{-\infty}^{0} \frac{x}{2\sigma} e^{\frac{-|x|}{\sigma}} dx \qquad (10)$$

since integrating over all values of $(-x)$ from $0 \to \infty$ is equivalent to integrating over all values of x from $-\infty \to 0$. Splitting the integral from (9) and applying the result from (10) yields

$$\mathbb{E}[X] = \int_{-\infty}^{0} \frac{x}{2\sigma} e^{\frac{-|x|}{\sigma}} dx + \int_0^{\infty} \frac{x}{2\sigma} e^{\frac{-|x|}{\sigma}} dx = \int_{-\infty}^{0} \frac{x}{2\sigma} e^{\frac{-|x|}{\sigma}} dx + \big(-\int_{-\infty}^{0} \frac{x}{2\sigma} e^{\frac{-|x|}{\sigma}} dx\big) = 0 \qquad (11)$$

Since the first moment of X is 0 regardless of σ, we cannot express the first moment in terms of sigma meaningfully, so we have to consider the second moment. Applying the definition of the second moment of a continuous random variable, we find

$$\mathbb{E}[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_{-\infty}^{\infty} \frac{x^2}{2\sigma} e^{\frac{-|x|}{\sigma}} dx \tag{12}$$

Note that, for all $x \in \mathbb{R}$,

$$\frac{(-x)^2}{2\sigma} e^{\frac{-|-x|}{\sigma}} = \frac{x^2}{2\sigma} e^{\frac{-|x|}{\sigma}}$$

so $\frac{x^2}{2\sigma} e^{\frac{-|x|}{\sigma}}$ is an even function, so we know

$$\int_0^{\infty} \frac{x^2}{2\sigma} e^{\frac{-|x|}{\sigma}} dx = \int_0^{\infty} \frac{(-x)^2}{2\sigma} e^{\frac{-|-x|}{\sigma}} dx = \int_{-\infty}^{0} \frac{x^2}{2\sigma} e^{\frac{-|x|}{\sigma}} dx \tag{13}$$

since integrating over all values of $(-x)$ from $0 \to \infty$ is equivalent to integrating over all values of x from $-\infty \to 0$. Splitting the integral from (12) and applying the result from (13) yields

$$\mathbb{E}[X^2] = \int_{-\infty}^{0} \frac{x^2}{2\sigma} e^{\frac{-|x|}{\sigma}} dx + \int_0^{\infty} \frac{x^2}{2\sigma} e^{\frac{-|x|}{\sigma}} dx = 2\int_0^{\infty} \frac{x^2}{2\sigma} e^{\frac{-|x|}{\sigma}} dx = \int_0^{\infty} \frac{x^2}{\sigma} e^{\frac{-|x|}{\sigma}} dx = \int_0^{\infty} \frac{x^2}{\sigma} e^{\frac{-x}{\sigma}} dx \tag{14}$$

with the last equality following from the fact that $|x| = x$ for all $x \in [0, \infty)$.
By definition, a Gamma distributed random variable Y with parameters α and β has PDF

$$f_Y(y) := \begin{cases} \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)} \text{ if } x > 0 \\ 0 \text{ otherwise.} \end{cases}$$

Thus, a Gamma distributed random variable Y with parameters $\alpha = 3$ and $\beta = \sigma$ would have PDF

$$f_Y(y) := \begin{cases} \frac{x^{3-1} e^{-\frac{x}{\sigma}}}{\sigma^3\Gamma(3)} \text{ if } x > 0 \\ 0 \text{ otherwise.} \end{cases} = \begin{cases} \frac{x^2 e^{-\frac{x}{\sigma}}}{\sigma^3\Gamma(3)} \text{ if } x > 0 \\ 0 \text{ otherwise.} \end{cases} \tag{15}$$

Comparing the integrand in (14) with the definition from (15), we find

$$\frac{x^2}{\sigma} e^{\frac{-x}{\sigma}} = \sigma^2\Gamma(3) f_Y(x) \tag{16}$$

Plugging the equality from (16) into (14) yields

$$\mathbb{E}[X^2] = \int_0^{\infty} \sigma^2\Gamma(3) f_Y(x)dx = \sigma^2\Gamma(3) \int_0^{\infty} f_Y(x)dx \tag{17}$$

Note, since $f_Y(x) = 0$ for all $x \leq 0$, we know

$$\int_{-\infty}^{\infty} f_Y(x)dx = \int_{-\infty}^{0} 0dx + \int_0^{\infty} f_Y(x)dx = \int_0^{\infty} f_Y(x)dx \tag{18}$$

Also, by the third axiom of probability and the definition of a continuous random variable, we have

$$\int_{-\infty}^{\infty} f_Y(x)dx = \mathbb{P}(\Omega) = 1 \tag{19}$$

where Ω is the sample space of the continuous random variable. Combining the rsults from (18) and (19) and plugging into (17) yields

$$\mathbb{E}[X^2] = \sigma^2\Gamma(3) \int_{-\infty}^{\infty} f_Y(x)dx = \sigma^2\Gamma(3) \cdot 1 = \sigma^2\Gamma(3) \tag{20}$$

We can use the fact that, for all $n \in \mathbb{N}$,

$$\Gamma(n+1) = n!$$

to quickly compute that

$$\Gamma(3) = \Gamma(2+1) = 2! = 2 \qquad (21)$$

Plugging the result from (21) into (20) yields

$$\mathbb{E}[X^2] = 2\sigma^2 \qquad (22)$$

Solving (22) for σ yields

$$\sigma = \sqrt{\frac{\mathbb{E}[X^2]}{2}} \qquad (23)$$

Let $X_1, \cdots, X_n$ be *i.i.d.* random variables with the same distribution as X. That is, for all $i \in \{1, ..., n\}$, X_i has PDF

$$f_{X_i}(x) := \frac{1}{2\sigma} e^{\frac{-|x|}{\sigma}}$$

for all $x \in \mathbb{R}$. Then the second sample moment of X is

$$M_2(X_1, ..., X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i^2 \qquad (24)$$

Substituting the second sample moment $M_2(X_1, ..., X_n)$ for X's second moment $\mathbb{E}[X^2]$ in (23) and applying the result from (24), we can define the method of moments estimator for σ to be

$$Z := \sqrt{\frac{M_2(X_1, ..., X_n)}{2}} = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{2}} = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{2n}} \qquad (25)$$

Thus, our method of moments estimator for σ is

$$Z := \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{2n}}$$

Exercise 9. Suppose you know that the following list of numbers is a random sample of size 20 from a Gaussian distribution with mean 1 and unknown variance $\sigma^2 > 0$.

$$2.0753 \quad 4.6678 \quad -3.5177 \quad 2.7243 \quad 1.6375 \quad -1.6154 \quad 0.1328 \quad 1.6852 \quad 8.1568 \quad 6.5389$$

$$-1.6998 \quad 7.0698 \quad 2.4508 \quad 0.8739 \quad 2.4295 \quad 0.5901 \quad 0.7517 \quad 3.9794 \quad 3.8181 \quad 3.8344.$$

- Using a method of moments estimator, estimate the value of σ^2 for this data. (Hint: Since the mean is 1, the variance σ^2 is equal to the second moment minus 1.)

  Solution.
  Define $X \sim Normal(1, \sigma^2)$ to be a Gaussian random variable with mean 1 and unknown variance $\sigma^2 > 0$. Let $X_1, ..., X_n$ be *i.i.d.* Gaussian random variables with mean 1 and the same unknown variance $\sigma^2 > 0$. We apply the hint to find that

  $$\sigma^2 = \mathbb{E}[X^2] - 1 \qquad (26)$$

  We can define the second sample moment of X to be

  $$M_2(X_1, ..., X_n) := \frac{1}{n} \sum_{i=1}^{n} X_i^2 \qquad (27)$$

Substituting the second sample moment $M_2(X_1, ..., X_n)$ for X's second moment $\mathbb{E}[X^2]$ into (26), we can define a method of moments estimator for σ^2 to be

$$Z := M_2(X_1, ..., X_n) - 1 = \big(\frac{1}{n}\sum_{i=1}^{n} X_i^2\big) - 1 \qquad (28)$$

Thus, our method of moments estimator for σ^2 is

$$Z := \big(\frac{1}{n}\sum_{i=1}^{n} X_i^2\big) - 1$$

Plugging in our sample data for $X_1, ..., X_n$, we find that our method of moments estimate for σ^2 for this data is

$$Z = \big(\frac{1}{20}\sum_{i=1}^{20} X_i^2\big) - 1 \approx 12.7452$$

- Denote your method of moments estimator for σ^2 as Z. Is Z unbiased?

Solution.
Claim: Z is an unbiased estimator for σ^2.
Proof: By the definition of an unbiased estimator for σ^2, it suffices to show that

$$\mathbb{E}[Z] = \sigma^2 \qquad (29)$$

Note that for all constants $a, b \in \mathbb{R}$, and for all random variables X,

$$\mathbb{E}[aX] = a\mathbb{E}[X] \qquad (30)$$

and

$$\mathbb{E}[X + b] = \mathbb{E}[X] + b \qquad (31)$$

Using (30) and (31) with our definition of Z from (28), we find

$$\mathbb{E}[Z] = \mathbb{E}[\big(\frac{1}{n}\sum_{i=1}^{n} X_i^2\big) - 1] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i^2] - 1 = \big(\frac{1}{n}\mathbb{E}[\sum_{i=1}^{n} X_i^2]\big) - 1 \qquad (32)$$

For all continuous random variables X and Y, we have

$$\begin{aligned}
\mathbb{E}[X+Y] = & \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{X,Y}(x,y)dxdy \\
= & \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf_{X,Y}(x,y)dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf_{X,Y}(x,y)dxdy \\
= & \int_{-\infty}^{\infty}x\int_{-\infty}^{\infty}f_{X,Y}(x,y)dydx + \int_{-\infty}^{\infty}y\int_{-\infty}^{\infty}f_{X,Y}(x,y)dxdy \\
= & \int_{-\infty}^{\infty}xf_X(x)dx + \int_{-\infty}^{\infty}yf_Y(y)dy = \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}$$

by the definitions of the marginals and expected value of continuous random variables. Similarly, for

all discrete random variables X and Y, we have

$$\begin{aligned}
\mathbb{E}[X+Y] = & \sum_{x\in\mathbb{R}}\sum_{y\in R}(x+y)\mathbb{P}(X=x,Y=y) \\
= & \sum_{x\in\mathbb{R}}\sum_{y\in\mathbb{R}}x\mathbb{P}(X=x,Y=y)+\sum_{x\in\mathbb{R}}\sum_{y\in\mathbb{R}}y\mathbb{P}(X=x,Y=y) \\
= & \sum_{x\in\mathbb{R}}x\sum_{y\in\mathbb{R}}\mathbb{P}(X=x,Y=y)+\sum_{y\in\mathbb{R}}y\sum_{x\in\mathbb{R}}\mathbb{P}(X=x,Y=y) \\
= & \sum_{x\in\mathbb{R}}x\mathbb{P}(X=x)+\sum_{y\in\mathbb{R}}y\mathbb{P}(Y=y)=\mathbb{E}[X]+\mathbb{E}[Y]
\end{aligned}$$

by the definitions of the marginals and expected value of discrete random variables. Therefore, for all random variables X and Y, we have

$$\mathbb{E}[X+Y]=\mathbb{E}[X]+\mathbb{E}[Y] \tag{33}$$

Applying the result from (33) to (32) $n-1$ times yields

$$\mathbb{E}[Z]=\big(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2]\big)-1 \tag{34}$$

Since we defined $X_1,...,X_n$ to be *i.i.d.*, we know

$$\mathbb{E}[X_1^2]=\cdots=\mathbb{E}[X_n^2]$$

Substituting $\mathbb{E}[X_1^2]$ for $\mathbb{E}[X_i^2]$ in (34) yields

$$\mathbb{E}[Z]=\big(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_1^2]\big)-1=\big(\frac{1}{n}\cdot n\mathbb{E}[X_1^2]\big)-1=\mathbb{E}[X_1^2]-1 \tag{35}$$

Note that we defined X_1 to be a Gaussian random variable with mean 1 and unknown variance $\sigma^2>0$. Note also that

$$Var(X_1)=\mathbb{E}[X_1^2]-(\mathbb{E}[X_1])^2 \implies \mathbb{E}[X_1^2]=Var(X_1)+(\mathbb{E}[X_1])^2$$

Since X_1 has mean $\mu_{X_1}=\mathbb{E}[X_1]=1$ and variance $\sigma^2>0$, we know

$$\mathbb{E}[X_1^2]=\sigma^2+1^2=\sigma^2+1 \tag{36}$$

Plugging the result from (36) into (35), we find

$$\mathbb{E}[Z]=\sigma^2+1-1=\sigma^2$$

By the definition of an unbiased estimator, this completes the proof that Z is unbiased for σ^2.

- We know for sure that $\sigma^2>0$. Is it possible that Z could take negative values? If so, then perhaps Z is not the best way to estimate σ^2.

Solution.
Claim: It is possible for Z to take negative values.
Proof: Assume we have a sample of size n in which $X_i<1$ for each $i\in\{1,...,n\}$. Note that this is certainly possible as $\mathbb{E}[X_i]=1$ and $\sigma^2>0$ implies that X_i is not constant. For such a sample, we have $X_i^2<1$ for all $i\in\{1,...,n\}$. Combining this inequality with (28) yields

$$Z:=\big(\frac{1}{n}\sum_{i=1}^{n}X_i^2\big)-1<\big(\frac{1}{n}\sum_{i=1}^{n}1\big)-1=\frac{n}{n}-1=0$$

Thus, such a possible sample yields a negative value for Z, so Z can take negative values. Since we are given $\sigma^2 > 0$, this suggests that Z is not the best estimator for σ^2, as Z can take on values which we know with certainty σ^2 cannot take.

- The Delta Method suggests that $1/Z$ could be a good estimate for $1/\sigma^2$. What estimate of $1/\sigma^2$ do you get from the data above? Is $\mathbb{E}\,|1/Z|$ finite? If not, then we cannot even compute the bias of this estimator. (Note that the distribution of Z should be closely related to a chi-squared distribution.) (Optional: if you use the fact that $\lim_{\varepsilon \to 0^+} \int_{\varepsilon < |t| < 1} \frac{1}{t} dt = 0$, then you should be able to estimate $\mathbb{E}(1/Z)$ as the number of samples n goes to infinity.)

  Solution.
  With the data above, we get an estimate for $\frac{1}{\sigma^2}$ of

$$\frac{1}{Z} \approx \frac{1}{12.7452} \approx 0.0785$$

  Similar to the previous proof that Z can take negative values, Z can also approach 0. Assume a sample of size n, where $X_i \in [1-\varepsilon, 1+\varepsilon]$ for all $i \in \{1, ..., n\}$ and for some $\varepsilon > 0$, which is entirely possible since $\mathbb{E}[X_i] = 1$ for all $i \in \{1, ..., n\}$ and each X_i is continuous. Then, as $\varepsilon \to 0^+$,

$$Z - \big(\frac{1}{n}\sum_{i=1}^{n} 1^2\big) - 1 \approx \big(\frac{1}{n}\sum_{i=1}^{n} 1\big) - 1 = \frac{n}{n} - 1 = 1 - 1 = 0$$

  Thus $Z \approx 0$ is a possibility, and Z is continuous, so we know Z can approach 0. As Z approaches 0, we have

$$\lim_{Z \to 0^+} \frac{1}{Z} = \infty, \qquad \lim_{Z \to 0^-} \frac{1}{Z} = -\infty$$

  Thus, it is possible for the random variable $\frac{1}{Z}$ to be infinite. The expected value of any random variable which can possibly take infinity as a value is not finite. Thus, we can conclude $\mathbb{E}[\frac{1}{Z}]$ is not finite.

- The Delta Method also suggests that Z^2 could be a good estimate for σ^4. What estimate of σ^4 do you get from the data above? Is Z^2 an unbiased estimate of σ^4?

  Solution.
  With the data above, we get an estimate for σ^4 of

$$Z^2 \approx (12.7452)^2 \approx 162.4397$$

  Claim: Z^2 is *not* an unbiased estimator for σ^4.
  Proof: Note that, by the definition of variance

$$Var(Z) = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2$$

  which implies

$$\mathbb{E}[Z^2] = Var(Z) + (\mathbb{E}[Z])^2 \qquad (37)$$

  We already computed that $\mathbb{E}[Z] = \sigma^2$, so we can quickly determine that

$$(\mathbb{E}[Z])^2 = (\sigma^2)^2 = \sigma^2\sigma^2 = \sigma^4 \qquad (38)$$

  We can apply the fact that, for all constants $a, b \in R$ and random variables X, we have

$$Var(aX + b) = a^2 Var(X)$$

to (28) to find

$$Var(Z) = Var((\frac{1}{n}\sum_{i=1}^{n} X_i^2) - 1) = \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i^2) \qquad (39)$$

Since $X_1, ..., X_n$ are *i.i.d.*, we know $X_1^2, ..., X_n^2$ are *i.i.d.*, so we can apply $Var(X + Y) = Var(X) + Var(Y)$ for all independent variables X, Y $n - 1$ times to find

$$Var(Z) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i^2) \qquad (40)$$

Also since $X_1^2, ..., X_n^2$ are *i.i.d.*, we know

$$Var(X_1^2) = \cdots = Var(X_n^2)$$

Substituting $Var(X_1^2)$ for $Var(X_i^2)$ for each $i \in \{1, ..., n\}$ in (40) yields

$$Var(Z) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_1)^2 = \frac{n}{n^2} Var(X_1^2) = \frac{Var(X_1^2)}{n} \qquad (41)$$

Note that

$$Var(X_1^2) = \mathbb{E}(X_1^4) - (\mathbb{E}[X_1^2])^2 \qquad (42)$$

by the definition of variance. Since

$$Var(X_1) = \mathbb{E}[X_1^2] - (\mathbb{E}[X_1])^2 \implies \mathbb{E}[X_1^2] = Var(X_1) + (\mathbb{E}[X_1])^2$$

and X_1 is a Gaussian random variable with mean $\mathbb{E}[X_1] = 1$ and variance $Var(X_1) = \sigma^2$, we can easily compute that

$$\mathbb{E}[X_1^2] = \sigma^2 + 1^2 = \sigma^2 + 1 \qquad (43)$$

To compute $\mathbb{E}[X_1^4]$, we define $T = X - 1$. Since, for any Gaussian random variable X and any constant $a \in \mathbb{R}$, $X + a$ is also a Gaussian, we know T is a Gaussian random variable. We can easily compute that

$$\mathbb{E}[T] = \mathbb{E}[X_1 - 1] = \mathbb{E}[X_1] - 1 = 1 - 1 = 0 \qquad (44)$$

and

$$Var(T) = Var(X_1 - 1) = Var(X_1) = \sigma^2 \qquad (45)$$

So T is a mean 0, variance σ^2 Gaussian random variable.
Claim: For any mean 0, variance σ^2 Gaussian random variable X and any continuous function g, we have

$$\mathbb{E}[g(X)X] = \sigma^2 \mathbb{E}[g'(X)]$$

Proof: By definition of a Gaussian random variable with mean 0 and variance σ^2, we know X has PDF

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \quad \forall x \in \mathbb{R}$$

Applying the definition of expected value, we find

$$\mathbb{E}[g(X)X] = \int_{-\infty}^{\infty} g(X) x f_X(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} g(X) x e^{-\frac{x^2}{2\sigma^2}} dx$$

Integrating by parts with $u = g(X)$, $du = g'(X)dx$, $dv = xe^{-\frac{x^2}{2\sigma^2}} dx$, $v = -\sigma^2 e^{-\frac{x^2}{2\sigma^2}}$ yields

$$\mathbb{E}[g(X)X] = \frac{1}{\sigma\sqrt{2\pi}} [-\sigma^2 g(X) e^{-\frac{x^2}{2\sigma^2}} |_{-\infty}^{\infty} + \sigma^2 \int_{-\infty}^{\infty} g'(X) e^{-\frac{x^2}{2\sigma^2}} dx]$$

Applying L'Hopital's Rule repeatedly to the leftmost term yields

$$-\sigma^2 g(X)e^{-\frac{x^2}{2\sigma^2}}|_{-\infty}^{\infty} = 0 - 0 = 0$$

so we are left with

$$E[g(X)X] = \frac{1}{\sigma\sqrt{2\pi}}\sigma^2 \int_{-\infty}^{\infty} g'(X)e^{-\frac{x^2}{2\sigma^2}}dx = \sigma^2 \int_{-\infty}^{\infty} g'(X)f_X(x)dx = \sigma^2\mathbb{E}[g'(X)]$$

with the last equality following by the definition of expected value. This completes the proof that $\mathbb{E}[g(X)X] = \sigma^2\mathbb{E}[g'(X)]$. We can now use this result since $T = X - 1$ is a mean 0, variance σ^2 Gaussian random variable. Note that

$$\mathbb{E}[X_1^4] = \mathbb{E}[(T-1)^4] = \mathbb{E}[T^4 - 4T^3 + 6T^2 - 4T + 1] \tag{46}$$

Applying linearity of expectation, we find

$$\mathbb{E}[X_1^4] = \mathbb{E}[T^4] - 4\mathbb{E}[T^3] + 6\mathbb{E}[T^2] - 4\mathbb{E}[T] + 1 \tag{47}$$

We can use the $\mathbb{E}[g(T)T] = \mathbb{E}[g'(T)]$ with $g(T) = T^3$ to find

$$\mathbb{E}[T^4] = \mathbb{E}[T^3T] = \sigma^2\mathbb{E}[3T^2] = 3\sigma^2\mathbb{E}[TT] = 3\sigma^4\mathbb{E}[1] = 3\sigma^4 \tag{48}$$

Similarly, with $g(T) = T^2$, we find

$$\mathbb{E}[T^3] = \mathbb{E}[T^2T] = \sigma^2\mathbb{E}[2T] = \sigma^2\mathbb{E}[2 \cdot T] = \sigma^4\mathbb{E}[0] = 0 \tag{49}$$

and with $g(T) = T$, we find

$$\mathbb{E}[T^2] = \mathbb{E}[TT] = \sigma^2\mathbb{E}[1] = \sigma^2 \tag{50}$$

Plugging (48), (49), (50), and (44) into (47), we find

$$\mathbb{E}[X_1^4] = 3\sigma^4 - 0 + 6\sigma^2 - 0 + 1 = 3\sigma^4 + 6\sigma^2 + 1 \tag{51}$$

Combining (51), (43), and (42) yields

$$Var(X_1^2) = \mathbb{E}[X_1^4] - (\mathbb{E}[X_1^2])^2 = 3\sigma^4 + 6\sigma^2 + 1 - (\sigma^2+1)^2 = 3\sigma^4 + 6\sigma^2 + 1 - \sigma^4 - 2\sigma^2 - 1 = 2\sigma^4 + 4\sigma^2 \tag{52}$$

Plugging (52) into (41) yields

$$Var(Z) = \frac{Var(X_1^2)}{n} = \frac{2\sigma^4 + 4\sigma^2}{n} \tag{53}$$

Plugging (53) and (38) into (37), we find

$$\mathbb{E}[Z^2] = Var(Z) + (\mathbb{E}[Z])^2 = \frac{2\sigma^4 + 4\sigma^2}{n} + \sigma^4 \tag{54}$$

Comparing (54) with σ^4, we find

$$\mathbb{E}[Z^2] = \frac{2\sigma^4 + 4\sigma^2}{n} + \sigma^4 \neq \sigma^4$$

By the definition of an unbiased estimator, this completes the proof that Z^2 is *not* an unbiased estimator for σ^4.

- Is Z^2 an asymptotically unbiased estimate of σ^4? That is, as the number of samples n goes to infinity, does $\mathbb{E}Z^2$ converge to σ^4?

  Solution. Claim: Z^2 is an *asymptotically* unbiased estimator of σ^4.
  Proof: It suffices to show that

$$\lim_{n\to\infty} \mathbb{E}[Z^2] = \sigma^4$$

  Using our equation for $\mathbb{E}[Z^2]$ from (54), we find

$$\lim_{n\to\infty} \mathbb{E}[Z^2] = \lim_{n\to\infty} (\frac{2\sigma^4 + 4\sigma^2}{n} + \sigma^4) = \lim_{n\to\infty} \frac{2\sigma^4 + 4\sigma^2}{n} + \lim_{n\to\infty} \sigma^4 \qquad (55)$$

  Since σ^4 doesn't depend on n, we know

$$\lim_{n\to\infty} \sigma^4 = \sigma^4 \qquad (56)$$

  We can directly compute that

$$\lim_{n\to\infty} \frac{2\sigma^4 + 4\sigma^2}{n} = 0 \qquad (57)$$

  Plugging (56) and (57) into (55) yields

$$\lim_{n\to\infty} \mathbb{E}[Z^2] = 0 + \sigma^4 = \sigma^4$$

  Thus, as the number of samples n goes to infinity, $\mathbb{E}[Z^2]$ converges to σ^4, so Z^2 is an asymptotically unbiased estimate of σ^4.

Exercise 10 (Conditional Expectation as a Random Variable)**.** Let $X, Y, Z\colon \Omega \to \mathbb{R}$ be discrete or continuous random variables. Let A be the range of Y. Define $g\colon A \to \mathbb{R}$ by $g(y) := \mathbb{E}(X|Y = y)$, for any $y \in A$. We then define the **conditional expectation** of X given Y, denoted $\mathbb{E}(X|Y)$, to be the random variable $g(Y)$.

(i) Let X, Y be random variables such that (X, Y) is uniformly distributed on the triangle $\{(x, y) \in \mathbb{R}^2\colon x \geq 0, y \geq 0, x + y \leq 1\}$. Show that

$$\mathbb{E}(X|Y) = \frac{1}{2}(1 - Y).$$

Solution.
Note: By the definition of conditional expectation, we have

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx \qquad (58)$$

By the definitions of the conditional PDF $f_{X|Y}(x|y)$ and the Y marginal PDF $f_Y(y)$, we have

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \qquad (59)$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx \qquad (60)$$

First, we have to solve for $f_{X,Y}(x, y)$. Since (X, Y) is uniformly distributed over $A = \{(x, y) \in \mathbb{R}^2 : x \geq 0, y \geq 0, x + y \leq 1\}$, we know the joint PDF

$$f_{X,Y}(x, y) = \begin{cases} a \text{ if } (x, y) \in A \\ 0 \text{ otherwise} \end{cases}$$

where $a \in \mathbb{R}$ is some unknown constant. Combining this with the third Axiom of probability that $\mathbb{P}(\Omega) = 1$, we have

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx = \int_0^1\int_0^{1-y} adxdy = a\int_0^1\int_0^{1-y} dxdy \tag{61}$$

Directly evaluating the right hand side of (61) yields

$$1 = a\int_0^1 x|_0^{1-y}dy = a\int_0^1 1-ydy = a[y - \frac{y^2}{2}]|_0^1 = a(1-\frac{1}{2}) = \frac{a}{2} \tag{62}$$

Note that (62) directly implies that $a = 2$, so we have

$$f_{X,Y}(x,y) = \begin{cases} 2 \text{ if } (x,y) \in A \\ 0 \text{ otherwise} \end{cases} \tag{63}$$

Plugging (63) into (60) yields

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_0^{1-y} 2dx = 2x|_0^{1-y} = 2(1-y)$$

so we know

$$f_Y(y) = \begin{cases} 2(1-y) \text{ if } 0 \le y \le 1 \\ 0 \text{ otherwise} \end{cases} \tag{64}$$

Plugging (63) and (64) into (59) yields

$$f_{X|Y}(x|y) = \frac{2}{2(1-y)} = \frac{1}{1-y}$$

so we know

$$f_{X|Y}(x|y) = \begin{cases} \frac{1}{1-y} \text{ if } (x,y) \in A \\ 0 \text{ if } 0 \le y \le 1, (x,y) \notin A \\ \text{undefined otherwise.} \end{cases} \tag{65}$$

Plugging (65) into (58) yields

$$\mathbb{E}[X|Y=y] = \int_0^{1-y} x\frac{1}{1-y}dx = \frac{1}{1-y}\int_0^{1-y} xdx = \frac{1}{1-y}\frac{x^2}{2}|_0^{1-y} = \frac{1}{1-y}\frac{(1-y)^2}{2} = \frac{1-y}{2} \tag{66}$$

for $0 \le y \le 1$. Also, for all $y < 0$ and for all $y > 1$, we have $f_Y(y) = 0$, so $f_{X|Y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$ is undefined. Similarly, for all $0 \le y \le 1$, $x > 1$ and for all $0 \le y \le 1$, $x < 0$, $f_{X,Y}(x,y) = 0$, so $f_{X|Y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = 0$. Thus, $\mathbb{E}[X|Y=y]$ is only defined and nonzero when $(x,y) \in A$, and $\mathbb{E}[X|Y=y] = \frac{1-y}{2}$ for all $(x,y) \in A$. Combining this with the definition of the conditional expectation of X given Y concludes the proof that

$$\mathbb{E}[X|Y] = \frac{1-Y}{2} = \frac{1}{2}(1-Y)$$

(ii) Prove the following version of the Total Expectation Theorem

$$\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X).$$

Solution.

First, we consider the case where X, Y are continuous random variables. Then, applying the definition of expected value of continuous random variables, and noting that $\mathbb{E}[X|Y]$ is a function of Y and not X, we find

$$\mathbb{E}[\mathbb{E}[X|Y]] = \int_{-\infty}^{\infty} \mathbb{E}[X|Y=y] f_Y(y) dy \tag{67}$$

Plugging $\int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx$ for $\mathbb{E}[X|Y=y]$ in (67) yields

$$\mathbb{E}[\mathbb{E}[X|Y]] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y}(x|y) f_Y(y) dx dy \tag{68}$$

Plugging (59) in for $f_{X|Y}(x|y)$ in (68) yields

$$\mathbb{E}[\mathbb{E}[X|Y]] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y) dx dy$$

Since none of the bounds depend on x or y, we can switch the order of integration in (69) to find

$$\mathbb{E}[\mathbb{E}[X|Y]] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X,Y}(x,y) dy dx = \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy dx \tag{70}$$

By definition of the X marginal, the inner integral

$$\int_{-\infty}^{\infty} f_{X,Y}(x,y) dy = f_X(x) \tag{71}$$

Plugging (71) into (70) yields

$$\mathbb{E}[\mathbb{E}[X|Y]] = \int_{-\infty}^{\infty} x f_X(x) dx \tag{72}$$

By the definition of expected value,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx \tag{73}$$

Plugging (73) into (72), we find

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X] \tag{74}$$

This completes the proof that $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$ for all continuous random variables X and Y.

Now, we will address the discrete case. Applying the definition of the expected value of discrete random variables, we find

$$\mathbb{E}[\mathbb{E}[X|Y]] = \sum_{y \in \mathbb{R}} \mathbb{E}[X|Y] \mathbb{P}(Y=y) \tag{75}$$

Plugging $\sum_{x \in \mathbb{R}} x \mathbb{P}(X=x|Y=y)$ in for $\mathbb{E}[X|Y]$ in (75) yields

$$\mathbb{E}[\mathbb{E}[X|Y]] = \sum_{y \in \mathbb{R}} \sum_{x \in \mathbb{R}} x \mathbb{P}(X=x|Y=y) \mathbb{P}(Y=y) \tag{76}$$

Note that, by the definition of conditional probability,

$$\mathbb{P}(X=x|Y=y) = \frac{\mathbb{P}(X=x, Y=y)}{\mathbb{P}(Y=y)} \implies \mathbb{P}(X=x|Y=y)\mathbb{P}(Y=y) = \mathbb{P}(X=x, Y=y) \tag{77}$$

Plugging the result from (77) into (76) yields

$$\mathbb{E}[\mathbb{E}[X|Y]] = \sum_{y\in\mathbb{R}}\sum_{x\in\mathbb{R}} x\mathbb{P}(X=x,Y=y) \qquad (78)$$

Switching the order of summations in (78) yields

$$\mathbb{E}[\mathbb{E}[X|Y]] = \sum_{x\in\mathbb{R}} x\sum_{y\in R} \mathbb{P}(X=x,Y=y) \qquad (79)$$

By definition of the X marginal of a discrete random variable X, we know

$$\mathbb{P}(X=x) = \sum_{y\in\mathbb{R}} \mathbb{P}(X=x,Y=y) \qquad (80)$$

Plugging (80) into (79) yields

$$\mathbb{E}[\mathbb{E}[X|Y]] = \sum_{x\in\mathbb{R}} x\mathbb{P}(X=x) \qquad (81)$$

By definition of the expected value of a discrete random variable X, we know

$$\mathbb{E}[X] = \sum_{x\in\mathbb{R}} x\mathbb{P}(X=x) \qquad (82)$$

Plugging (82) into (81) yields

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X] \qquad (83)$$

which completes the proof that $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$ for all discrete random variables X and Y. The combination of (83) and (74) completes the general proof that $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.

- (Optional) If X is a random variable, and if $f(t) := \mathbb{E}(X-t)^2$, $t\in\mathbb{R}$, then the function $f\colon \mathbb{R}\to\mathbb{R}$ is uniquely minimized when $t = \mathbb{E}X$. A similar minimizing property holds for conditional expectation. Let $h\colon \mathbb{R}\to\mathbb{R}$. Show that the quantity $\mathbb{E}(X-h(Y))^2$ is minimized among all functions $h\colon \mathbb{R}\to\mathbb{R}$ when $h(Y) = \mathbb{E}(X|Y)$. (Hint: use the previous item.)

  Solution.
  It suffices to show that

$$\mathbb{E}[(X-h(Y))^2] \geq \mathbb{E}[(X-\mathbb{E}[X|Y])^2]$$

for all $h:\mathbb{R}\to\mathbb{R}$. We can add and subtract $\mathbb{E}[X|Y]$ to the interior of $\mathbb{E}[(X-h(Y))^2]$ to find

$$\mathbb{E}[(X-h(Y))^2] = \mathbb{E}[(X-\mathbb{E}[X|Y]+\mathbb{E}[X|Y]-h(Y))^2] \qquad (84)$$

Grouping the interior of (84) into $(X-\mathbb{E}[X|Y])+(\mathbb{E}[X|Y]-h(Y))$ and expanding yields

$$\mathbb{E}[(X-h(Y))^2] = \mathbb{E}[(X-\mathbb{E}[X|Y])^2 + 2(X-\mathbb{E}[X|Y])(\mathbb{E}[X|Y]-h(Y)) + (\mathbb{E}[X|Y]-h(Y))^2] \qquad (85)$$

Applying linearity of expectation to (85) yields

$$\mathbb{E}[(X-h(Y))^2] = \mathbb{E}[(X-\mathbb{E}[X|Y])^2]+\mathbb{E}[(\mathbb{E}[X|Y]-h(Y))^2]+2\mathbb{E}[(X-\mathbb{E}[X|Y])(\mathbb{E}[X|Y]-h(Y))] \qquad (86)$$

By the Total Expectation Theorem, we know

$$\mathbb{E}[(X-\mathbb{E}[X|Y])(\mathbb{E}[X|Y]-h(Y))] = \mathbb{E}[\mathbb{E}[(X-\mathbb{E}[X|Y])(\mathbb{E}[X|Y]-h(Y))|Y]] \qquad (87)$$

Since $\mathbb{E}[X|Y]$ and $h(Y)$ both already depend on Y, $\mathbb{E}[\mathbb{E}[X|Y]|Y] = \mathbb{E}[X|Y]$ and $\mathbb{E}[h(Y)|Y] = \mathbb{E}[h(Y)]$. Plugging these results and (87) into (86) yields

$$\begin{aligned}\mathbb{E}[(X-h(Y))^2] = &\quad \mathbb{E}[(X-\mathbb{E}[X|Y])^2] + \mathbb{E}[(\mathbb{E}[X|Y]-h(Y))^2] + 2\mathbb{E}[(\mathbb{E}[X|Y]-\mathbb{E}[X|Y](\mathbb{E}[X|Y]-h(Y)]] \\ = &\quad \mathbb{E}[(X-\mathbb{E}[X|Y])^2] + \mathbb{E}[(\mathbb{E}[X|Y]-h(Y))^2] + 2\mathbb{E}[0*(\mathbb{E}[X|Y]-h(Y))] \\ = &\quad \mathbb{E}[(X-\mathbb{E}[X|Y])^2] + \mathbb{E}[(\mathbb{E}[X|Y]-h(Y))^2] + 2\mathbb{E}[0] \\ = &\quad \mathbb{E}[(X-\mathbb{E}[X|Y])^2] + \mathbb{E}[(\mathbb{E}[X|Y]-h(Y))^2] \qquad (88)\end{aligned}$$

Since $a^2 \geq 0$ for all $a \in R$, we know $(\mathbb{E}[X|Y]-h(Y))^2 \geq 0$, so we know

$$\mathbb{E}[(\mathbb{E}[X|Y]-h(Y))^2] \geq 0 \qquad (89)$$

Combining (89) with (88) yields

$$\mathbb{E}[(X-h(Y))^2] = \mathbb{E}[(X-\mathbb{E}[X|Y])^2]+\mathbb{E}[(\mathbb{E}[X|Y]-h(Y))^2] \geq \mathbb{E}[(X-\mathbb{E}[X|Y])^2]+0 = \mathbb{E}[(X-\mathbb{E}[X|Y])^2] \quad (90)$$

Thus, we have shown that, for any $h : \mathbb{R} \to \mathbb{R}$

$$\mathbb{E}[(X-h(Y))^2] \geq \mathbb{E}[(X-\mathbb{E}[X|Y])^2]$$

which completes the proof that $h(Y) = \mathbb{E}[X|Y]$ minimizes $\mathbb{E}[(X-h(Y))^2]$ among all functions $h : \mathbb{R} \to \mathbb{R}$.

(iv) Show the following

$$\mathbb{E}(X|X) = X.$$
$$\mathbb{E}(X+Y|Z) = \mathbb{E}(X|Z) + \mathbb{E}(Y|Z).$$

Solution.
First, we will show $\mathbb{E}[X|X] = X$. Since we are given X, and $X = X$ tautologically, we know that

$$f_{X,X}(y,x) = \begin{cases} f_X(x) \text{ if } x = y \\ 0 \text{ otherwise.} \end{cases} \qquad (91)$$

Plugging (91) into (59), we find

$$f_{X|X=x}(y|x) = \begin{cases} \frac{f_X(x)}{f_X(x)} = 1 \text{ if } y = x \\ 0 \text{ otherwise.} \end{cases} \qquad (92)$$

Note that, for all constants $a \in \mathbb{R}$, we have

$$f_a(x) := \begin{cases} 1 \text{ if } x = a \\ 0 \text{ otherwise} \end{cases} \qquad (93)$$

Comparing (93) and (92), we find that $X|X = x$ is a constant with value x. Since $\mathbb{E}[a] = a$ for all constants $a \in \mathbb{R}$, we know

$$\mathbb{E}[X|X = x] = (X|X = x) = x \qquad (94)$$

By the definition of the conditional expectation of X given X, the conclusion that

$$\mathbb{E}[X|X] = X$$

follows, which completes the proof.
Note: The discrete case follows by exactly the same logic, simply replacing all PDFs f with corresponding PMFs p.

Now, we will show that $\mathbb{E}[(X+Y)|Z] = \mathbb{E}[X|Z] + \mathbb{E}[Y|Z]$. First, consider the case where X, Y, Z are discrete random variables. Then by the definition of the discrete conditional expectation of $X+Y$ given $Z = z$, we have

$$\mathbb{E}[(X+Y)|Z=z] = \sum_{t=x+y, x,y\in\mathbb{R}} t\mathbb{P}(X+Y=t|Z=z) = \sum_{x\in\mathbb{R}}\sum_{y\in\mathbb{R}}(x+y)\mathbb{P}(X=x, Y=y|Z=z) \tag{95}$$

Distributing $(x+y)$ and splitting the resulting sum yields

$$\begin{aligned}\mathbb{E}[(X+Y)|Z=z] = &\sum_{x\in R}\sum_{y\in\mathbb{R}} x\mathbb{P}(X=x, Y=y|Z=z) + \sum_{x\in R}\sum_{y\in\mathbb{R}} y\mathbb{P}(X=x, Y=y|Z=z) \\ = &\sum_{x\in\mathbb{R}} x\sum_{y\in\mathbb{R}}\mathbb{P}(X=x, Y=y|Z=Z) + \sum_{y\in\mathbb{R}} y\sum_{x\in\mathbb{R}}\mathbb{P}(X=x, Y=y|Z=Z)\end{aligned} \tag{96}$$

By the definition of the $X|Z$ marginal of a discrete random variable $X|Z=z$, we have

$$\mathbb{P}(X=x|Z=z) = \sum_{y\in\mathbb{R}}\mathbb{P}(X=x, Y=y|Z=z) \tag{97}$$

and similarly

$$\mathbb{P}(Y=y|Z=Z) = \sum_{x\in\mathbb{R}}\mathbb{P}(X=x, Y=y|Z=z) \tag{98}$$

Plugging (97) and (98) into (96) yields

$$\mathbb{E}[(X+Y)|Z=z] = \sum_{x\in\mathbb{R}} x\mathbb{P}(X=x|Z=z) + \sum_{y\in\mathbb{R}} y\mathbb{P}(Y=y|Z=z) \tag{99}$$

By the definitions of the conditional expectations of discrete random variables $X|Z=z$ and $Y|Z=z$, respectively, we know

$$\mathbb{E}[X|Z=z] = \sum_{x\in\mathbb{R}} x\mathbb{P}(X=x|Z=z) \tag{100}$$

and

$$\mathbb{E}[Y|Z=z] = \sum_{y\in\mathbb{R}} y\mathbb{P}(Y=y|Z=z) \tag{101}$$

Plugging (100) and (101) into (99), we find

$$\mathbb{E}[(X+Y)|Z=z] = \mathbb{E}[X|Z=z] + \mathbb{E}[Y|Z=z]$$

The conclusion that

$$\mathbb{E}[(X+Y)|Z] = \mathbb{E}[X|Z] + \mathbb{E}[Y|Z]$$

follows by the definition of conditional expectation for all discrete random variables X, Y, Z.
For continuous random variables, we follow a similar proof. This time, by the definition of the continuous conditional expectation of $X+Y$ given $Z=z$, we have

$$\mathbb{E}[(X+Y)|Z=z] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{X,Y|Z}(x,y|z)dxdy \tag{102}$$

Distributing $(x+y)$ to the integrand of (102) and splitting the integral yields

$$\mathbb{E}[(X+Y)|Z=z] = \int_{-\infty}^{\infty} x\int_{-\infty}^{\infty} f_{X,Y|Z}(x,y|z)dydx + \int_{-\infty}^{\infty} y\int_{-\infty}^{\infty} f_{X,Y|Z}(x,y|z)dxdy \tag{103}$$

By the definitions of the $X|Z$ and $Y|Z$ marginals of the continuous random variables $X|Z=z$ and $Y|Z=z$, we know

$$f_{X|Z}(x|z) = \int_{-\infty}^{\infty} f_{X,Y|Z}(x,y|z)dy \qquad (104)$$

and

$$f_{Y|Z}(y|z) = \int_{-\infty}^{\infty} f_{X,Y|Z}(x,y|z)dx \qquad (105)$$

Plugging (105) and (104) into (103) yields

$$\mathbb{E}[(X+Y)|Z=z] = \int_{-\infty}^{\infty} x f_{X|Z}(x|z)dx + \int_{-\infty}^{\infty} y f_{Y|Z}(y|z)dy \qquad (106)$$

By the definitions of the conditional expectations of the continuous random variables $X|Z=z$ and $Y|Z=z$, we have

$$\mathbb{E}[X|Z=z] = \int_{-\infty}^{\infty} x f_{X|Z}(x|z)dx \qquad (107)$$

and

$$\mathbb{E}[Y|Z=z] = \int_{-\infty}^{\infty} y f_{Y|Z}(y|z)dy \qquad (108)$$

Combining (107) and (108) with (106) yields

$$\mathbb{E}[(X+Y)|Z=z] = \mathbb{E}[X|Z=z] + \mathbb{E}[Y|Z=z]$$

The conclusion that

$$\mathbb{E}[(X+Y)|Z] = \mathbb{E}[X|Z] + \mathbb{E}[Y|Z]$$

follows by the definition of the conditional expectation of $X+Y$ given Z for all continuous random variables X, Y, Z. This combines with the previous proof for the discrete case to complete the proof that

$$\mathbb{E}[(X+Y)|Z] = \mathbb{E}[X|Z] + \mathbb{E}[Y|Z]$$

for any random variables X, Y, Z.

(v) If Z is independent of X and Y, show that

$$\mathbb{E}(X|Y,Z) = \mathbb{E}(X|Y).$$

(Here $\mathbb{E}(X|Y,Z)$ is notation for $\mathbb{E}(X|(Y,Z))$ where (Y,Z) is interpreted as a random vector, so that X is conditioned on the random vector (Y,Z).)

Solution.
First, we will prove the statement for discrete random variables X, Y, Z. By the definition of the conditional expectation of the discrete random variable X given $Y=y, Z=z$, we know

$$\mathbb{E}[X|Y=y, Z=z] = \sum_{x\in\mathbb{R}} x\mathbb{P}(X=x|Y=y, Z=z) \qquad (109)$$

Applying the definition of conditional probability to (109) yields

$$\mathbb{E}[X|Y=y, Z=z] = \sum_{x\in\mathbb{R}} x\frac{\mathbb{P}(X=x, Y=y, Z=z)}{\mathbb{P}(Y=y, Z=z)} \qquad (110)$$

Since Z is independent of X and Y, we know

$$\mathbb{P}(X=x, Y=y, Z=z) = \mathbb{P}(X=x, Y=y)\mathbb{P}(Z=z) \qquad (111)$$

and

$$\mathbb{P}(Y = y, Z = z) = \mathbb{P}(Y = y)\mathbb{P}(Z = z) \tag{112}$$

Plugging (111) and (112) into (110) yields

$$\begin{aligned}
\mathbb{E}[X|Y = y, Z = z] = & \sum_{x\in\mathbb{R}} x\frac{\mathbb{P}(X = x, Y = y)\mathbb{P}(Z = z)}{\mathbb{P}(Y = y)\mathbb{P}(Z = z)} \\
= & \sum_{x\in\mathbb{R}} x\frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} \\
= & \sum_{x\in\mathbb{R}} x\mathbb{P}(X = x|Y = y) \qquad (113)
\end{aligned}$$

By the definition of the conditional expectation of the discrete random variable $X|Y = y$, we know that

$$\mathbb{E}[X|Y = y] = \sum_{x\in\mathbb{R}} x\mathbb{P}(X = x|Y = y) \tag{114}$$

Plugging (114) into (113) yields

$$\mathbb{E}[X|Y = y, Z = z] = \mathbb{E}[X|Y = y]$$

By the definition of the conditional expectation of X given Y, Z, this implies

$$\mathbb{E}[X|Y, Z] = \mathbb{E}[X|Y] \tag{115}$$

for all discrete random variables X, Y, Z s.t. Z is independent of X and Y.
Now, we can complete a similar proof for the continuous case. By the definition of the conditional expectation of the continuous random variable X given $Y = y, Z = z$, we know

$$\mathbb{E}[X|Y = y, Z = z] = \int_{-\infty}^{\infty} xf_{X|Y,Z}(x|y, z)dx = \int_{-\infty}^{\infty} x\frac{f_{X,Y,Z}(x, y, z)}{f_{Y,Z}(y, z)}dx \tag{116}$$

Since Z is independent of X and Y, we know

$$f_{X,Y,Z}(x, y, z) = f_{X,Y}(x, y)f_Z(z) \tag{117}$$

and

$$f_{Y,Z}(y, z) = f_Y(y)f_Z(z) \tag{118}$$

Plugging (118) and (117) into (116) yields

$$\begin{aligned}
\mathbb{E}[X|Y = y, Z = z] = & \int_{-\infty}^{\infty} x\frac{f_{X,Y}(x, y)f_Z(z)}{f_Y(y)f_Z(z)}dx \\
= & \int_{-\infty}^{\infty} x\frac{f_{X,Y}(x, y)}{f_Y(y)}dx \\
= & \int_{-\infty}^{\infty} xf_{X|Y}(x|y)dx \qquad (119)
\end{aligned}$$

By the definition of the conditional expectation of the continuous random variable $X|Y = y$, we know

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} xf_{X|Y}(x|y)dx \tag{120}$$

Plugging (120) into (119) yields

$$\mathbb{E}[X|Y = y, Z = z] = \mathbb{E}[X|Y = y]$$

The conclusion that

$$\mathbb{E}[X|Y,Z] = \mathbb{E}[X|Y]$$

follows by the definition of the conditional expectation of X given Y, Z for all continuous random variables X, Y, Z s.t. Z is independent of X and Y. This combines with the previous proof for discrete X, Y, Z to complete the proof that

$$\mathbb{E}[X|Y,Z] = \mathbb{E}[X|Y]$$

for all random variables X, Y, Z s.t. Z is independent of X and Y.

Exercise 11 (Sunspot Data, Version 2)**.** This exercise deals with sunspot data from the following files (the same data appears in different formats)

txt file csv (excel) file

These files are taken from http://www.sidc.be/silso/datafiles#total

To work with this data, e.g. in Matlab you can use the command

```
x=importdata('SN_d_tot_V2.0.txt')
```

to import the .txt file.

The format of the data is as follows.

- Columns 1-3: Gregorian calendar date (Year, Month, then Day)
- Column 4: Date in fraction of year
- Column 5: Daily total number of sunspots observed on the sun. A value of -1 indicates that no number is available for that day (missing value).
- Column 6: Daily standard deviation of the input sunspot numbers from individual stations.
- Column 7: Number of observations used to compute the daily value.
- Column 8: Definitive/provisional indicator. A blank indicates that the value is definitive. A '*' symbol indicates that the value is still provisional and is subject to a possible revision (Usually the last 3 to 6 months)

In a previous Exercise, we examined the number of sunspots U_t versus time t, where the units of t in the data are in integers divided by 365 (or by 365.25),. You should have observed that the sunspots had a roughly 11-year periodicity. To make this more precise, we will use an estimator that checks for frequencies present in the data.

Denote $i := \sqrt{-1}$. For any real number r, consider the following estimator

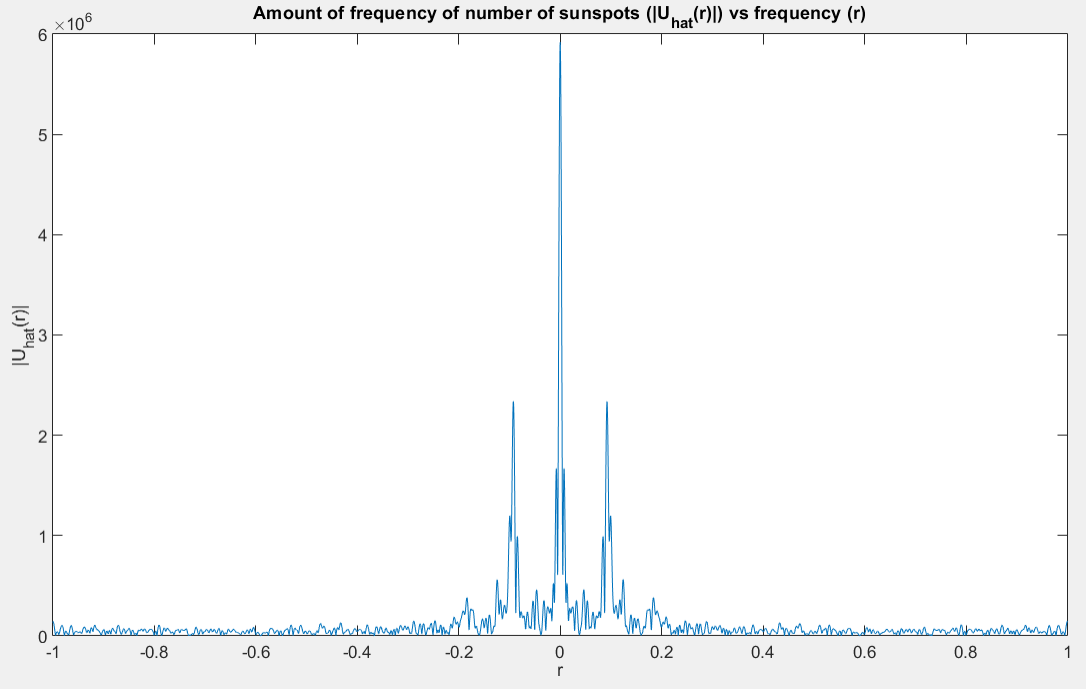
$$\widehat{U}(r) := \sum_{t \in \mathbb{Z}/365} U_t e^{2\pi i t r}.$$

This estimator measures the "amount" of frequency r that the number of sunspots has. (As usual $\mathbb{Z}$ denotes the set of integers.)

Plot $|\widehat{U}(r)|$ versus r, where $r \in [-1, 1]$. Do you observe any large absolute values of $\widehat{U}(r)$ for any values of r near $1/11$?

You should observe some large values of $\widehat{U}(r)$ when r takes the values: $.0842, .0921$, and $.0995$, corresponding to frequencies of $11.87, 10.858$, and 10.05, respectively. This large signal should correspond to $r \in [.08, .105]$ (and to $r \in [-.105, -.08]$).

Solution. Plotting $|\hat{U}(r)|$ versus r, where $r \in [-1, 1]$, yields the following graph:



As expected, there are large absolute values of $|\hat{U}(r)|$ around $r = \pm\frac{1}{11}$. Also as expected, the values of $|\hat{U}(r)|$ exhibit symmetrically large magnitudes over $r \in [0.08, 0.105]$ and $r \in [-.105, -0.08]$. Thus, our graph of $|\hat{U}(r)|$ versus r demonstrates expected behavior based on the roughly 11-year periodicity of U_t versus t.

Exercise 12. Let $\theta \in \mathbb{R}$ be an unknown parameter. Consider the density

$$f_\theta(x) := \begin{cases} e^{-(x-\theta)}, & \text{if } x \geq \theta \\ 0, & \text{if } x < \theta. \end{cases}$$

Suppose $X_1, \ldots, X_n$ is a random sample of size n, such that X_i has density f_θ for all $1 \leq i \leq n$.

Show that $X_{(1)} = \min_{1 \leq i \leq n} X_i$ is a sufficient statistic for θ.

Solution.
First, note that $X_{(1)}$ is a sufficient statistic for $\theta \iff$ the distribution of $(X_1, ..., X_n)|X_{(1)} = a$ does *not* depend on θ. Applying the definition of conditional density, we find

$$f_{X_1,...,X_n|X_{(1)}=a}(x_1, ..., x_n|a) = \frac{f_{X_1,...,X_n,X_{(1)}}(X_1 = x_1, ..., X_n = x_n, X_{(1)} = a)}{f_{X_{(1)}}(a)} \tag{121}$$

We can assume $a = \min_{1 \leq i \leq n} x_i$ so that

$$f_{X_1,...,X_n|X_{(1)}=a}(x_1, ..., x_n|a) = \frac{f_{X_1,...,X_n,X_{(1)}}(x_1, ..., x_n, \min_{1 \leq i \leq n} x_i)}{f_{X_{(1)}}(a)} \tag{122}$$

Since $\min_{1 \leq i \leq n} x_i$ is a function of $x_1, ..., x_n$, its value is entirely determined by $x_1, ..., x_n$, so we know

$$f_{X_1,...,X_n,X_{(1)}}(x_1, ..., x_n, \min_{1 \leq i \leq n} x_i) = f_{X_1,...,X_n}(x_1, ..., x_n) \tag{123}$$

This follows from $A \subseteq B \implies A \cap B = A$. Plugging (123) into (122) yields

$$f_{X_1,...,X_n|X_{(1)}=a}(x_1, ..., x_n|a) = \frac{f_{X_1,...,X_n}(x_1, ..., x_n)}{f_{X_{(1)}}(a)} \tag{124}$$

Since $X_1, ..., X_n$ are *i.i.d.* with density $f_\theta(x)$, we can directly compute that

$$\begin{aligned} f_{X_1,...,X_n}(x_1,...,x_n) = & \quad f_{X_1}(x_1)\cdot\cdots\cdot f_{X_n}(x_n) = f_\theta(x_1)\cdots f_\theta(x_n) \\ = & \quad 1_{X_1\geq\theta}(x_1)e^{-(x_1-\theta)}\cdot\cdots\cdot 1_{X_n\geq\theta}(x_n)e^{-(x_n-\theta)} \\ = & \quad 1_{X_1,...,X_n\geq\theta}(x_1,...,x_n)e^{-(x_1-\theta)+\cdots+-(x_n-\theta)} \\ = & \quad 1_{X_1,...,X_n\geq\theta}(x_1,...,x_n)e^{(-(x_1+\cdots+x_n)+n\theta)} \\ = & \quad 1_{X_1,...,X_n\geq\theta}(x_1,...,x_n)e^{-(x_1+\cdots+x_n)}e^{n\theta} \qquad (125) \end{aligned}$$

for all $x_1, ..., x_n \in \mathbb{R}^n$.
Also, we can use the fact that if X is a continuous random variable with density f_X and cumulative distribution function F_X, then for any $1 \leq j \leq n$, $F_{X_{(j)}}$ has density

$$f_{X_{(j)}}(x) := \frac{n!}{(j-1)!(n-j)!}f_X(x)(F_X(x))^{j-1}(1-F_X(x))^{n-j}, \qquad \forall\, x \in \mathbb{R}. \quad (126)$$

which was given in Exercise 4 of Homework 2, to compute $f_{X_{(1)}}(a)$. First, we need to compute $F_\theta(x)$. We directly find

$$F_\theta(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_\theta(t)dt = \int_\theta^x e^{-(t-\theta)}dt = -e^{-(t-\theta)}|_\theta^x = -e^{-(x-\theta)} + e^{-(\theta-\theta)} = 1 - e^{-(x-\theta)} \quad (127)$$

for all $\theta \leq x \leq \infty$, and $F_\theta(x) = 0$ otherwise. Plugging (127) and the given $f_\theta(x)$ into (126) yields

$$\begin{aligned} f_{X_{(1)}}(x) := & \quad \begin{cases} \frac{n!}{0!(n-1)!}e^{-(x-\theta)}(1-(1-e^{-(x-\theta)}))^{n-1} \text{ if } x \geq \theta \\ 0 \text{ otherwise.} \end{cases} \\ = & \quad \begin{cases} \frac{n!}{(n-1)!}e^{-(x-\theta)}[e^{-(x-\theta)}]^{n-1} \text{ if } x \geq \theta \\ 0 \text{ otherwise.} \end{cases} \\ = & \quad \begin{cases} n[e^{-(x-\theta)}]^n = ne^{-n(x-\theta)} = ne^{-nx+n\theta} \text{ if } x \geq \theta \\ 0 \text{ otherwise} \end{cases} \\ = & \quad \begin{cases} ne^{-nx}e^{n\theta} \text{ if } x \geq \theta \\ 0 \text{ otherwise.} \end{cases} \\ = & \quad 1_{X_{(1)}\geq\theta}(x)ne^{-nx}e^{n\theta} \\ = & \quad 1_{X_1,...,X_n\geq\theta}(x_1,...,x_n) \qquad (128) \end{aligned}$$

with the last equality following from the fact that $X_{(1)} \geq \theta \iff X_1, ..., X_n \geq \theta$.
Plugging (128) and (125) into (124) yields

$$f_{X_1,...,X_n|X_{(1)}=a}(x_1,...,x_n|a) = \frac{1_{X_1,...,X_n\geq\theta}(x_1,...,x_n)e^{-(x_1+\cdots+x_n)}e^{n\theta}}{1_{X_1,...,X_n\geq\theta}(x_1,...,x_n)ne^{-na}e^{n\theta}} = \frac{e^{-(x_1+\cdots+x_n)}}{ne^{-na}} \qquad (129)$$

Note that the right hand side of (129) expresses $f_{X_1,...,X_n|X_{(1)}=a}(x_1,...,x_n|a)$ in a way that does *not* depend on θ. Therefore, (129) implies that the distribution of $X_1, ..., X_n$ given $X_{(1)} = a$ is *not* dependent on θ, which completes the proof that $X_{(1)}$ is a sufficient statistic for θ.
Note: We could also apply the Factorization Theorem. Let

$$h(x_1,...,x_n) = e^{-(x_1+\cdots+x_n)}$$

and

$$g_\theta(a) = g_\theta(\min_{1\leq i\leq n} x_i) = 1_{X_{(1)}\geq\theta}(a)e^{n\theta}$$

Then $h(x_1, ..., x_n)$ depends only on our sample (*not* on θ), and $g_\theta(a)$ is a function of our statistic $X_{(1)} = \min_{1 \le i \le n} X_i$ that does depend on θ. Comparing $h(x_1, ..., x_n) \cdot g_\theta(a)$ with (125) yields

$$\begin{aligned} f_{X_1,...,X_n}(x_1, ..., x_n) = & \quad 1_{X_1,...,X_n \ge \theta}(x_1, ..., x_n) e^{-(x_1+\cdots+x_n)} e^{n\theta} \\ = & \quad e^{-(x_1+\cdots+x_n)} \cdot 1_{X_1,...,X_n \ge \theta}(x_1, ..., x_n) e^{n\theta} \\ = & \quad e^{-(x_1+\cdots+x_n)} \cdot 1_{X_{(1)} \ge \theta}(a) e^{n\theta} \\ = & \quad h(x_1, ..., x_n) \cdot g_\theta(a) \qquad (130) \end{aligned}$$

Since the equality from (130) holds for all $(x_1, ..., x_n) \in \mathbb{R}^n$ except a set of measure 0, the conclusion that $X_{(1)} = \min_{1 \le i \le n} X_i$ is a sufficient statistic for θ follows by the Factorization Theorem. This completes the alternative proof that $X_{(1)}$ is sufficient for θ.

Assignment 4

Mathematical Statistics 408 Steven Heilman

Please provide complete and well-written solutions to the following exercises.
Due October 12, 12PM noon PST, to be uploaded as a single PDF document to Gradescope.

Homework 4 - Emerson Kahle

Exercise 13. Let $X_1, \ldots, X_n$ be a random sample of size n from a Poisson distribution with unknown parameter $\lambda > 0$. (So, $\mathbb{P}(X_1 = k) = e^{-\lambda}\lambda^k/k!$ for all integers $k \ge 0$.)
Let Y be the estimator $Y = 1_{\{X_1=0\}}$. Suppose we want to estimate $e^{-\lambda}$.

- Find a method of moments estimator for $e^{-\lambda}$. Is this estimator consistent?
- Show that Y is unbiased for $e^{-\lambda}$.
- Show that $\sum_{i=1}^n X_i$ is sufficient for $e^{-\lambda}$.
- Compute $W_n := \mathbb{E}_\lambda(Y \mid \sum_{i=1}^n X_i)$, as in the Rao-Blackwell Theorem.
- As $n \to \infty$, does W_n converge in any sense? If so, what does it converge to? Does this mean that $W_1, W_2, \ldots$ is consistent?

Solution.

(a) By definition of the expected value of a discrete random variable, we have

$$\mathbb{E}[X_1] = \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \qquad (1)$$

Since the first term in this sum, $0 * e^{-\lambda}\frac{\lambda^0}{0!} = 0$, we can rewrite (1) as

$$\mathbb{E}[X_1] = 0 + \sum_{k=1}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \qquad (2)$$

Moving the constant $\lambda e^{-\lambda}$ outside the sum and simplifying $k \cdot \frac{1}{k!}$ in (2) yields

$$\mathbb{E}[X_1] = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \qquad (3)$$

By the power series definition for the exponential function, we know

$$e^x := \sum_{i=0}^{\infty} \frac{x^i}{i!} \qquad (4)$$

Plugging the definition from (4) into (3) yields

$$\mathbb{E}[X_1] = \lambda e^{-\lambda} e^{\lambda} = \lambda \qquad (5)$$

Negating then applying the exponential function to both sides of (5) yields

$$e^{-\lambda} = e^{-\mathbb{E}[X_1]} \qquad (6)$$

Note that (6) expresses $e^{-\lambda}$ as a function of the first moment $\mathbb{E}[X_1] = \mu_1$ Define $M_1(X_1, ..., X_n)$ to be the first sample moment. That is,

$$M_1(X_1, ..., X_n) := \frac{1}{n} \sum_{i=1}^{n} X_i \qquad (7)$$

Substituting M_1 for $\mu_1 = \mathbb{E}[X_1]$ in (6) yields

$$Z_n := e^{-M_1(X_1,..,X_n)} = e^{-(\frac{\sum_{i=1}^{n} X_i}{n})} \qquad (8)$$

Thus, a method of moments estimator for $e^{-\lambda}$ is

$$Z_n := e^{-(\frac{\sum_{i=1}^{n} X_i}{n})}$$

Assuming

$$\mathbb{E}[X_1] = \lambda < \infty \qquad (9)$$

the Weak Law of Large Numbers guarantees that

$$\mathbb{P}(|\frac{X_1 + \cdots + X_n}{n} - \lambda| > \varepsilon) = 0$$

for all $\varepsilon > 0$ as $n \to \infty$. That is, the first sample moment (sample mean) is consistent for $\mathbb{E}[X_1] = \lambda$ (the population mean). So the sequence $\overline{X}_1, \overline{X}_2, ...$ converges in probability $\mathbb{E}[X_1] = \lambda$ (where $\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$). By Exercise 2.36 (from the Notes), we know for any continuous function $f : \mathbb{R} \to \mathbb{R}$, the sequence $f(\overline{X}_1), f(\overline{X}_2), ...$ converges in probability to $f(\mathbb{E}[X_1]) = f(\lambda)$. Thus, if we let $f(x) := e^{-x}$, which is continuous across all of $\mathbb{R}$, we know the sequence $e^{-\overline{X}_1}, e^{-\overline{X}_2}, ...$ converges in probability to $e^{-\mathbb{E}[X_1]} = e^{-\lambda}$. That is,

$$\lim_{n\to\infty} \mathbb{P}(|e^{-\overline{X}_n} - e^{-\lambda}| > \varepsilon) = 0 \qquad (10)$$

for all $\varepsilon > 0$. We defined our method of moments estimator Z_n in (8) to be

$$Z_n := e^{-(\frac{\sum_{i=1}^{n} X_i}{n})} = e^{-\overline{X}_n} \qquad (11)$$

The combination of (10) and (11) implies

$$\lim_{n\to\infty} \mathbb{P}(|Z_n - e^{-\lambda}| > \varepsilon) = 0 \qquad (12)$$

for all $\varepsilon > 0$. By the definition of consistency, since the sequence $Z_1, Z_2, ...$ converges in probability to the constant value $f(\mathbb{E}[X_1]) = f(\lambda) = e^{-\lambda}$, for all $0 < \lambda < \infty$, we know our sequence of method of moments estimators $Z_1, Z_2, ...$ is consistent for $e^{-\lambda}$.

(b) Since Y is defined to be an indicator function, we know $Y \in \{0, 1\}$. By the definition of the expected value of a discrete random variable, we have

$$\mathbb{E}[Y] = 0 \cdot \mathbb{P}(Y = 0) + 1 \cdot \mathbb{P}(Y = 1) = \mathbb{P}(Y = 1) \tag{13}$$

Since $Y = 1_{X_1=0} = \begin{cases} 1 \text{ if } X_1 = 0 \\ 0 \text{ otherwise} \end{cases}$, (13) yields

$$\mathbb{E}[Y] = \mathbb{P}(Y = 1) = \mathbb{P}(X_1 = 0) \tag{14}$$

Since X_1 is a Poisson distributed random variable with parameter $\lambda > 0$, we know

$$\mathbb{P}(X_1 = 0) = e^{-\lambda}\frac{\lambda^0}{0!} = e^{-\lambda} \tag{15}$$

Plugging (15) into (14) yields

$$\mathbb{E}[Y] = e^{-\lambda} \tag{16}$$

By the definition of an unbiased estimator, this completes the proof that Y is unbiased for $e^{-\lambda}$.

(c) Since $X_1, ..., X_n$ are *i.i.d.* Poisson distributed random variables with parameter $\lambda > 0$, we know each X_i has PMF

$$f_{X_i}(x_i) := \mathbb{P}(X_i = x_i) = e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} \tag{17}$$

for all $i \in \{1, ..., n\}$. Since $X_1, ..., X_n$ are independent, we know the joint PMF of $X_1, ..., X_n$ is just the product of the individual PMFs of $X_1, ..., X_n$. That is,

$$f_{X_1,...,X_n}(x_1, ..., x_n) := \mathbb{P}(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) \tag{18}$$

Plugging the result from (17) into (18) and simplifying yields

$$f_{X_1,...,X_n}(x_1, ..., x_n) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = (e^{-\lambda}\frac{\lambda^{x_1}}{x_1!}) \cdot \cdots \cdot (e^{-\lambda}\frac{\lambda^{x_n}}{x_n!}) = e^{-\lambda n}\frac{\lambda^{x_1+\cdots+x_n}}{x_1! \cdot \cdots \cdot x_n!} \tag{19}$$

Noting that $\lambda = -ln(e^{-\lambda})$, we can rewrite the joint PMF of $X_1, ..., X_n$ from (19) as

$$f_{X_1,...,X_n}(x_1, ..., x_n) = (e^{-\lambda})^n(-ln(e^{-\lambda}))^{x_1+\cdots+x_n}\frac{1}{x_1! \cdot \cdots \cdot x_n!} \tag{20}$$

Now, let $g_{e^{-\lambda}}(a) := (e^{-\lambda})^n(-ln(e^{-\lambda}))^a$, let $h(x_1, ..., x_n) = \frac{1}{x_1! \cdot \cdots \cdot x_n!}$ and let $t(X_1, ..., X_n) := \sum_{i=1}^{n} X_i = X_1 + \cdots + X_n$. Note that g depends on $e^{-\lambda}$ while h depends only on the sample values that $X_1, ..., X_n$ take. We can rewrite (20) as

$$f_{X_1,...,X_n}(x_1, ..., x_n) = g_{e^{-\lambda}}(t(x_1, ..., x_n))h(x_1, ..., x_n) \tag{21}$$

Thus, we can write the joint PMF of $X_1, ..., X_n$ as the product of $g_{e^{-\lambda}}(t(x_1, ..., x_n))$ and $h(x_1, ..., x_n)$. Since $g_{e^{-\lambda}}(t(x_1, ..., x_n))$ is a function of our statistic t which *does* depend on $e^{-\lambda}$, and $h(x_1, ..., x_n)$ is a function of our sample $X_1, ..., X_n$ which does *not* depend on $e^{-\lambda}$, the Factorization Theorem guarantees that $t(X_1, .., X_n) := X_1 + \cdots + X_n = \sum_{i=1}^{n} X_i$ is sufficient for $e^{-\lambda}$. This completes the proof that $\sum_{i=1}^{n} X_i$ is a sufficient statistic for $e^{-\lambda}$.

(d) Since Y being an indicator function implies $Y \in \{0, 1\}$, we know the conditional expectation of Y given $\sum_{i=1}^{n} X_i = k$ is

$$\mathbb{E}_\lambda[Y|\sum_{i=1}^{n} X_i = k] = 0 \cdot \mathbb{P}(Y = 0|\sum_{i=1}^{n} X_i = k) + 1 \cdot \mathbb{P}(Y = 1|\sum_{i=1}^{n} X_i = k) = \mathbb{P}(Y = 1|\sum_{i=1}^{n} X_i = k) \tag{22}$$

Applying the definition of conditional probability to (22) yields

$$\mathbb{E}_\lambda[Y|\sum_{i=1}^{n} X_i = k] = \frac{\mathbb{P}(Y=1, \sum_{i=1}^{n} X_i = k)}{\mathbb{P}(\sum_{i=1}^{n} X_i = k)} \tag{23}$$

As noted in part (b), $Y = 1 \iff X_1 = 0$, so we can rewrite the numerator from (23) as

$$\mathbb{P}(Y=1, \sum_{i=1}^{n} X_i = k) = \mathbb{P}(X_1 = 0, \sum_{i=1}^{n} X_i = k) \tag{24}$$

Moreover, $X_1 = 0 \implies \sum_{i=1}^{n} X_i = \sum_{i=2}^{n} X_i$, so we know $(X_1 = 0, \sum_{i=1}^{n} X_i = k) \iff (X_1 = 0, \sum_{i=2}^{n} X_i = k)$. This implies that

$$\mathbb{P}(X_1 = 0, \sum_{i=1}^{n} X_i = k) = \mathbb{P}(X_1 = 0, \sum_{i=2}^{n} X_i = k) \tag{25}$$

Plugging the result from (25) into (24) yields

$$\mathbb{P}(Y=1, \sum_{i=1}^{n} X_i = k) = \mathbb{P}(X_1 = 0, \sum_{i=2}^{n} X_i = k) \tag{26}$$

Since $X_1, ..., X_n$ are independent, we know X_1 is independent of $\sum_{i=2}^{n} X_i$, so we know

$$\mathbb{P}(X_1 = x, \sum_{i=2}^{n} X_i = k) = \mathbb{P}(X_1 = x)\mathbb{P}(\sum_{i=2}^{n} X_i = k) \tag{27}$$

Plugging (26) into (23) and applying the result from (27) yields

$$\mathbb{E}_\lambda[Y|\sum_{i=1}^{n} X_i = k] = \frac{\mathbb{P}(X_1 = 0)\mathbb{P}(\sum_{i=2}^{n} X_i = k)}{\mathbb{P}(\sum_{i=1}^{n} X_i = k)} \tag{28}$$

Now, note that for any 2 independent Poisson random variables X, Y with parameters $\lambda_1, \lambda_2 > 0$, we have $X + Y = k \iff X = i, Y = k - i$ for some $i \in \{0, ..., k\}$, so

$$\mathbb{P}(X + Y = k) = \mathbb{P}(\bigcup_{i=0}^{k} (X = i, Y = k - i)) \tag{29}$$

For all $i \neq j$, $i, j \in \{0, ..., k\}$, $(X = i, Y = k - i)$ and $(X = j, Y = k - j)$ are mutually disjoint events. Applying the second axiom of probability to (29) yields

$$\mathbb{P}(X + Y = k) = \sum_{i=0}^{k} \mathbb{P}(X = i, Y = k - i) \tag{30}$$

Since X and Y are independent, we know $\mathbb{P}(X = i, Y = k - i) = \mathbb{P}(X = i)\mathbb{P}(Y = k - i)$ for all $i \in \{0, ..., k\}$. Applying this result to (30) yields

$$\mathbb{P}(X + Y = k) = \sum_{i=0}^{k} \mathbb{P}(X = i)\mathbb{P}(Y = k - i) \tag{31}$$

Since $X \sim Poisson(\lambda_1)$ and $Y \sim Poisson(\lambda_2)$, we know

$$\mathbb{P}(X = k) = e^{-\lambda_1} \frac{\lambda_1^k}{k!} \quad \mathbb{P}(Y = k) = e^{-\lambda_2} \frac{\lambda_2^k}{k!} \tag{32}$$

Plugging the result from (32) into (31) yields

$$\mathbb{P}(X+Y=k)=\sum_{i=0}^{k}e^{-\lambda_1}\frac{\lambda_1^i}{i!}e^{-\lambda_2}\frac{\lambda_2^{k-i}}{(k-i)!}=e^{-(\lambda_1+\lambda_2)}\sum_{i=0}^{k}\frac{1}{i!(k-i)!}\lambda_1^i\lambda_2^{k-i} \tag{33}$$

Noting that $\binom{k}{i}:=\frac{k!}{i!(k-i)!}$ so $\frac{1}{i!(k-i)!}=\frac{1}{k!}\binom{k}{i}$, and plugging this result into (33) yields

$$\mathbb{P}(X+Y=k)=e^{-(\lambda_1+\lambda_2)}\sum_{i=0}^{k}\frac{1}{k!}\binom{k}{i}\lambda_1^i\lambda_2^{k-i}=e^{-(\lambda_1+\lambda_2)}\frac{1}{k!}\sum_{i=0}^{k}\binom{k}{i}\lambda_1^i\lambda_2^{k-i} \tag{34}$$

Applying the Binomial Theorem to (34) yields

$$\mathbb{P}(X+Y=k)e^{-(\lambda_1+\lambda_2)}\frac{1}{k!}\sum_{i=0}^{k}\binom{k}{i}\lambda_1^i\lambda_2^{k-i}=e^{-(\lambda_1+\lambda_2)}\frac{1}{k!}(\lambda_1+\lambda_2)^k=e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^k}{k!} \tag{35}$$

Note that the PMF of $X+Y$ in (35) is just the PMF of a Poisson distributed random variable with parameter $\lambda_1+\lambda_2>0$. Since (35) holds for all $k\in\mathbb{Z}$ such that $k\geq 0$, we know that, for any two independent Poisson distributed random variables X and Y with parameters $\lambda_1,\lambda_2>0$, the random variable $X+Y$ is Poisson distributed with parameter $\lambda_1+\lambda_2>0$. Using the fact that $X_1,...,X_n$ are *i.i.d.* Poisson distributed random variables with parameter $\lambda>0$, and applying the previous result $n-1$ times, we find

$$\mathbb{P}(X_1+\cdots+X_n=k)=e^{-(\sum_{i=1}^{n}\lambda)}\frac{(\sum_{i=1}^{n}\lambda)^k}{k!}=e^{-n\lambda}\frac{(n\lambda)^k}{k!} \tag{36}$$

Since the PMF for $X_1+\cdots+X_n$ from (36) is just the PMF of a Poisson random variable with parameter $n\lambda>0$, and (36) holds for all $k\in\mathbb{Z}$ such that $k\geq 0$, we know $\sum_{i=1}^{n}X_i$ is a Poisson distributed random variable with parameter $n\lambda$. Similarly, $\sum_{i=2}^{n}X_i$ is a Poisson random variable with parameter $(n-1)\lambda$. That is,

$$\mathbb{P}(\sum_{i=2}^{n}X_i=k)=e^{-(n-1)\lambda}\frac{((n-1)\lambda)^k}{k!} \tag{37}$$

for all $k\in\mathbb{Z}$ such that $k\geq 0$.
Plugging the results from (36) and (37) into (28) yields

$$\mathbb{E}_\lambda[Y|\sum_{i=1}^{n}X_i=k]=\frac{e^{-\lambda}\frac{\lambda^0}{0!}e^{-(n-1)\lambda}\frac{((n-1)\lambda)^k}{k!}}{e^{-n\lambda}\frac{(n\lambda)^k}{k!}}=\frac{e^{-\lambda-(n-1)\lambda}\frac{((n-1)\lambda)^k}{k!}}{e^{-n\lambda}\frac{(n\lambda)^k}{k!}}=\frac{e^{-n\lambda}\frac{((n-1)\lambda)^k}{k!}}{e^{-n\lambda}\frac{(n\lambda)^k}{k!}} \tag{38}$$

Cancelling like terms to simplify (38) yields

$$\mathbb{E}_\lambda[Y|\sum_{i=1}^{n}X_i=k]=\frac{((n-1)\lambda)^k}{(n\lambda)^k}=(\frac{(n-1)\lambda}{n\lambda})^k=(\frac{n-1}{n})^k=(\frac{n}{n}-\frac{1}{n})^k=(1-\frac{1}{n})^k=:g(k) \tag{39}$$

where $k=\sum_{i=1}^{n}X_i$. By the definition of Conditional Expectation as a random variable, we can substitute $\sum_{i=1}^{n}X_i$ for k in (39) to find

$$W_n:=\mathbb{E}_\lambda[Y|\sum_{i=1}^{n}X_i]=(1-\frac{1}{n})^{\sum_{i=1}^{n}X_i} \tag{40}$$

Thus, the closed form expression for W_n is

$$W_n=(1-\frac{1}{n})^{\sum_{i=1}^{n}X_i}$$

This completes the computation of W_n.

(e) *Claim:* As $n \to \infty$, W_n converges in probability to $e^{-\lambda}$, and $W_1, W_2, ...$ is consistent for $e^{-\lambda}$.
Proof. We need to show

$$\lim_{n\to\infty} \mathbb{P}(|W_n - e^{-\lambda}| > \varepsilon) = 0 \qquad (*)$$

for all $\varepsilon > 0$. First, note that

$$\sum_{i=1}^{n} X_i = X_1 + \cdots + X_n = n\overline{X}_n \qquad (41)$$

Plugging the result from (41) into (40) yields

$$W_n = (1 - \frac{1}{n})^{n\overline{X}_n} = ((1 - \frac{1}{n})^n)^{\overline{X}_n} \qquad (42)$$

By Lemma 1.28 (from the Notes), we know that, for any sequence $\lambda_1, \lambda_2, ... > 0$ and number $\lambda^* > 0$ such that $\lim_{n\to\infty} \lambda_n = \lambda^*$, we have

$$\lim_{n\to\infty} (1 - \frac{\lambda_n}{n})^n = e^{-\lambda^*} \qquad (43)$$

Thus, for $\lambda_1, \lambda_2, ... = 1, 1, ...$ such that $\lim_{n\to\infty} \lambda_n = \lambda_n = 1 = \lambda^*$, (43) implies that

$$\lim_{n\to\infty} (1 - \frac{\lambda_n}{n})^n = \lim_{n\to\infty} (1 - \frac{1}{n})^n = e^{-\lambda^*} = e^{-1} \qquad (44)$$

Writing W_n as in (42) and applying the result from (44) to (*) yields

$$\lim_{n\to\infty} (\mathbb{P}(|W_n - e^{-\lambda}| > \varepsilon) = \lim_{n\to\infty} \mathbb{P}(|(e^{-1})^{\overline{X}_n} - e^{-\lambda}| > \varepsilon) = \lim_{n\to\infty} \mathbb{P}(|e^{-\overline{X}_n} - e^{-\lambda}| > \varepsilon) \qquad (45)$$

Comparing (45) with (10), we find that

$$\lim_{n\to\infty} (\mathbb{P}(|W_n - e^{-\lambda}| > \varepsilon) = \lim_{n\to\infty} \mathbb{P}(|e^{-\overline{X}_n} - e^{-\lambda}| > \varepsilon) = 0 \qquad (46)$$

for all $\varepsilon > 0$.

Just like in part (a), this once again follows from a combination of three things:
(i) The Weak Law of Large Numbers guarantees that $\overline{X}_n$ converges in probability to $\mathbb{E}[X_1] = \lambda$;
(ii) Exercise 2.36 (from the Notes) guarantees that, for any sequence $Y_1, Y_2, ...$ that converges in probability to a random variable U and any continuous function $f : \mathbb{R} \to \mathbb{R}$, $f(Y_1), f(Y_2), ...$ converges in probability to $f(U)$;
(iii) $f(x) = e^{-x}$ is a continuous real-valued function from $\mathbb{R} \to \mathbb{R}$.

From (46), we see that W_n converges in probability to $e^{-\lambda}$ as $n \to \infty$. By the definition of consistency, since this holds for all $0 < \lambda < \infty$, we know the sequence $W_1, W_2, ...$ is consistent for $e^{-\lambda}$. This completes the proof that W_n converges in probability to $e^{-\lambda}$ as $n \to \infty$, and $W_1, W_2, ...$ is consistent for $e^{-\lambda}$.

Exercise 14. Let $X_1, \ldots, X_n$ be a random sample of size n from the uniform distribution on $[0, \theta]$ where $\theta > 0$ is unknown.

On a previous homework, we showed that

$$X_{(n)} = \max_{1\le i\le n} X_i$$

is a sufficient statistic for θ.

- Show that $2X_1$ is an unbiased estimator of θ.
- Compute $W := \mathbb{E}_\theta(2X_1 \,|\, X_{(n)})$, as in the Rao-Blackwell Theorem. (Hint: with probability $1/n$, $X_1 = X_{(n)}$. And with probability $1 - 1/n$, $X_1 < X_{(n)}$, and if additionally $X_{(n)} = x$, then X_1 is uniform on $(0, x)$.) Using whatever method you wish, show that W is unbiased for θ.
- A method of moments estimator for θ is $2\frac{1}{n}\sum_{i=1}^n X_i$. Compute

$$\mathbb{E}_\theta\Big(2\frac{1}{n}\sum_{i=1}^n X_i \,\Big|\, X_{(n)}\Big).$$

Solution.

(a) Since $\mathbb{E}[aX] = a\mathbb{E}[X]$ for all random variables X and constants $a \in \mathbb{R}$, we know

$$\mathbb{E}[2X_1] = 2\mathbb{E}[X_1] \qquad (47)$$

Since X_1 is uniformly distributed on $[0, \theta]$, we know X_1 has PDF

$$f_{X_1}(x) := \begin{cases} \frac{1}{\theta} \text{ if } x \in [0, \theta] \\ 0 \text{ otherwise.} \end{cases} \qquad (48)$$

Thus, we can directly compute that

$$\mathbb{E}[X_1] = \int_{-\infty}^{\infty} x f_{X_1}(x) dx = \int_0^\theta \frac{x}{\theta} dx = \frac{1}{\theta}[\frac{x^2}{2}\Big|_0^\theta] = \frac{1}{\theta}(\frac{\theta^2}{2}) = \frac{\theta}{2} \qquad (49)$$

Plugging the result from (49) into (47) yields

$$\mathbb{E}[2X_1] = 2\frac{\theta}{2} = \theta \qquad (50)$$

By the definition of an unbiased estimator, since $\mathbb{E}[2X_1] = \theta$ for all $0 < \theta < \infty$, this completes the proof that $2X_1$ is unbiased for θ.

(b) Since $\mathbb{E}[aX|Y] = a\mathbb{E}[X|Y]$ for all random variables X, Y and constants $a \in \mathbb{R}$, we have

$$\mathbb{E}[2X_1|X_{(n)} = k] = 2\mathbb{E}[X_1|X_{(n)} = k] \qquad (51)$$

Following the hint, if $X_{(n)} := \max_{1 \le i \le n} X_i = k$, then

$$P(X_1 = X_{(n)}) = \frac{1}{n} \qquad (52)$$

Since $(X_{(n)} = X_1)$ and $(X_{(n)} \neq X_1)$ are mutually exclusive, we can rewrite $\mathbb{E}[X_1|X_{(n)} = k]$ as

$$\begin{aligned} \mathbb{E}[X_1|X_{(n)} = k] = \quad & \mathbb{E}[X_1|X_{(n)} = k, X_1 = X_{(n)}]\mathbb{P}(X_1 = X_{(n)}) \\ + \quad & \mathbb{E}[X_1|X_{(n)} = k, X_1 \neq X_{(n)}]\mathbb{P}(X_1 \neq X_{(n)}) \end{aligned} \qquad (53)$$

This is analogous to the identity that, for any random variable X and event $A \subseteq \Omega$, $\mathbb{E}[X] = \mathbb{E}[X|A]\mathbb{P}(A) + \mathbb{E}[X|A^c]\mathbb{P}(A^c)$, which follows from the Law of Total Probability. If $X_1 = X_{(n)}$ and $X_{(n)} = k$, we know $X_1 = k$, so $X_1|X_{(n)} = k, X_1 = X_{(n)}$ is the constant random variable of value k. That is, $\mathbb{P}(X_1 = k|X_{(n)} = k, X_1 = X_{(n)}) = 1$. This implies

$$\mathbb{E}[X_1|X_{(n)} = k, X_1 = X_{(n)}] = k \qquad (54)$$

Also, since $(X_{(n)} = k) \implies X_1 \in (0, k]$, we know $(X_{(n)} = k, X_1 \neq X_{(n)}) \implies 0 \leq X_1 < k$. Since X_1 is uniformly distributed on $[0, \theta]$, restricting the upper bound of X_1 to $X_1 < k$ means that X_1 is uniformly distributed on $(0, k)$. Thus, given $X_{(n)} = k, X_1 \neq X_{(n)}$, we know

$$f_{X_1|X_{(n)}=k, X_1 \neq X_{(n)}}(x|k) = \begin{cases} \frac{1}{k} \text{ if } x \in (0, k) \\ 0 \text{ otherwise.} \end{cases} \tag{55}$$

The results from (54) and (55) motivate the need to split $\mathbb{E}[X_1|X_{(n)} = k]$ as in (53). When $X_{(n)} = X_1$, which occurs with probability $\frac{1}{n}$, X_1 is a constant random variable, so it takes exactly one value with nonzero probability, and is thus discrete. However, when $X_{(n)} \neq X_1$, which occurs with probability $1 - \frac{1}{n} = \frac{n-1}{n}$, X_1 is a continuous uniform random variable. By the definition of the expected value of a continuous random variable, we need a summation (with a single term) to compute the $\mathbb{E}[X_1|X_{(n)} = k]$ when $X_{(n)} = X_1$. However, by the definition of the expected value of a continuous random variable, we need an to integrate $f_{X_1|X_{(n)}=k, X_1 \neq X_{(n)}}(x|k)$ over $(0, k)$. Thus, we cannot compute $\mathbb{E}[X_1|X_{(n)} = k]$ without splitting the expectation to consider both the case when $X_{(n)} = X_1$ and the case when $X_{(n)} \neq X_1$.
Now, we can apply the definition of the conditional expectation of a continuous random variable, along with the result from (55), to find

$$\mathbb{E}[X_1|X_{(n)} = k, X_1 < X_{(1)}] = \int_{-\infty}^{\infty} x f_{X_1|X_{(n)}=k, X_1 \neq X_{(n)}}(x|k)dx = \int_0^k x \frac{1}{k} dx = \frac{1}{k}[\frac{x^2}{2}\Big|_0^k] = \frac{1}{k}\frac{k^2}{2} = \frac{k}{2} \tag{56}$$

Also, since $X_1 < X_{(n)} \iff X_1 \neq X_{(n)}$ (by definition of X_n as the maximum of $X_1, ..., X_n$), we know

$$\mathbb{P}(X_1 < X_{(n)}) = \mathbb{P}(X_1 \neq X_{(n)}) = 1 - \mathbb{P}(X_1 = X_{(n)}) \tag{57}$$

Plugging (52) into (57) yields

$$\mathbb{P}(X_1 < X_{(n)}) = 1 - \frac{1}{n} = \frac{n-1}{n}$$

Plugging (52), (54), (56), and (57) into (52) yields

$$\mathbb{E}[X_1|X_{(n)} = k] = k\frac{1}{n} + \frac{k}{2}\frac{n-1}{n} \tag{58}$$

Plugging (58) into (51) yields

$$\mathbb{E}[2X_1|X_{(n)} = k] = 2(k\frac{1}{n} + \frac{k}{2}\frac{n-1}{n}) = \frac{2k}{n} + \frac{k(n-1)}{n} = \frac{k(2+n-1)}{n} = \frac{n+1}{n}k =: g(k) \tag{59}$$

By the definition of conditional expectation as a random variable, since we assumed $X_{(n)} = k$, we can substitute $X_{(n)}$ for k in (59) to find

$$W := \mathbb{E}_\theta[2X_1|X_{(n)}] = \frac{n+1}{n}X_{(n)} \tag{60}$$

From Example 4.7 (from the Notes), we know $(1 + \frac{1}{n})X_{(n)} = \frac{n+1}{n}X_{(n)}$ is unbiased for θ. That is,

$$\mathbb{E}[\frac{n+1}{n}X_{(n)}] = \theta \tag{61}$$

Since $W = \frac{n+1}{n}X_{(n)}$ by (60), we know

$$\mathbb{E}[W] = \mathbb{E}[\frac{n+1}{n}X_{(n)}] = \theta \tag{61}$$

By the definition of an unbiased estimator for θ, this completes the proof that $W := \mathbb{E}_\theta[2X_1|X_{(n)}]$ is unbiased for θ.

We can also show that W is unbiased via the Law of Iterated Expectation (also known as the Law of Total Expectation), which states that for any two random variables X and Y where $\mathbb{E}[X]$ is well-defined,

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X] \qquad (62)$$

Since $\mathbb{E}[2X_1] = \theta$ is well-defined by part (a), we can apply (62) to find

$$\mathbb{E}[W] := \mathbb{E}[\mathbb{E}_\theta[2X_1|X_{(n)}]] = \mathbb{E}_\theta[2X_1] = \theta \qquad (63)$$

with the last equality following from equation (50) in part (a). By the definition of an unbiased estimator, this completes the alternative proof that $W := \mathbb{E}_\theta[2X_1|X_{(n)}]$ is unbiased for θ.

(c) *Claim:*

$$\mathbb{E}_\theta[2\frac{1}{n}\sum_{i=1}^{n} X_i|X_{(n)}] = \frac{n+1}{n}X_{(n)}$$

Proof. Using the fact that n is a real-valued constant and $\mathbb{E}[aX|Y] = a\mathbb{E}[X|Y]$ for all constants $a \in \mathbb{R}$, we have

$$\mathbb{E}_\theta[2\frac{1}{n}\sum_{i=1}^{n} X_i|X_{(n)} = k] = \frac{2}{n}\mathbb{E}_\theta[\sum_{i=1}^{n} X_i|X_{(n)} = k] \quad \forall k \geq 0 \qquad (64)$$

Note that $\mathbb{P}(X_{(n)} < 0) = 0$, so we only need to consider $X_{(n)}$ taking non-negative values k. By definition, $X_{(n)} \in \{X_1, ..., X_n\}$, so we know $\exists i\{1, ..., n\}$ such that $X_i = X_{(n)}$. Given that $X_{(n)} = k$, we have $X_i = k$. Also, we can safely assume that $X_j \neq X_{(n)} = k$ for all $j \in \{1, ..., n\}$ such that $j \neq i$ because

$$\begin{aligned}\mathbb{P}(X_j = X_{(n)}) = \quad & \mathbb{P}(X_j = X_i) = \int\int_{\{(x,y)\in\mathbb{R}^2|x=y\}} f_{X_j,X_i}(x,y)dxdy \\ = \quad & \int_{y=-\infty}^{\infty}\int_{x=y}^{y} f_{X_j,X_i}(x,y)dxdy = \int_{-\infty}^{\infty} 0dy = 0 \qquad (65)\end{aligned}$$

with the second to last equality following because $\int_a^a f dx = 0$ for all functions f. By (65), given $X_{(n)} = k$, we know $\exists i^* \in \{1, ..., n\}$ such that $X_{i^*} = X_{(n)} = k$ and $X_j \neq X_{(n)}$ for all $j \in \{1, ..., n\}$ such that $j \neq i^*$. By the definition of $X_{(n)}$ as the maximum of $X_1, ..., X_n$, we know $X_j < X_{i^*} = k$ for all $j \in \{1, ..., n\}$ such that $j \neq i^*$. Since $X_1, ..., X_n$ are i.i.d. uniform random variables on $[0, \theta]$, we know $\mathbb{P}(X_a < 0) = 0$ for all $a \in \{1, ..., n\}$, so $X_j \geq 0$ for all $j \in \{1, ..., n\}$ such that $j \neq i^*$. So we know $(X_{i^*}|X_{(n)} = k)$ is the constant random variable with $\mathbb{P}(X_{i^*} = k|X_{(n)} = k) = 1$, and $J := \{(X_j|j \in \{1, ..., n\}, j \neq i^*)|X_{(n)} = k\}$ is a set of $n-1$ *i.i.d.* random variables uniformly distributed on $(0, k)$. Define $Y_1, ..., Y_{n-1}$ such that $\{Y_1, ..., Y_{n-1}\} = J$. That is, $Y_1, ..., Y_{n-1}$ are $n-1$ i.i.d random variables uniformly distributed on $(0, k)$. Now, we can rewrite the interior of the expectation from (64) as

$$(\sum_{i=1}^{n} X_i|X_{(n)} = k) = (X_{i^*}|X_{(n)} = k) + \sum_{j=1}^{n-1} Y_j \qquad (66)$$

Plugging (66) into (64) and applying Exercise 4.14, we find

$$\mathbb{E}_\theta[2\frac{1}{n}\sum_{i=1}^{n} X_i|X_{(n)} = k] = \frac{2}{n}\mathbb{E}_\theta[(X_{i^*}|X_{(n)} = k) + \sum_{j=1}^{n-1} Y_j] = \frac{2}{n}\mathbb{E}_\theta[(X_{i^*}|X_{(n)} = k)] + \frac{2}{n}\mathbb{E}_\theta[\sum_{j=1}^{n-1} Y_j] \qquad (67)$$

Since $\mathbb{P}(X_{i^*} = k|X_{(n)} = k) = 1$, we know

$$\mathbb{E}[(X_{i^*}|X_{(n)} = k] = k \qquad (68)$$

Since $Y_1, ..., Y_{n-1}$ are identically distributed, we know

$$\mathbb{E}_\theta[\sum_{j=1}^{n-1} Y_j] = \mathbb{E}[Y_1] + \cdots + \mathbb{E}[Y_{n-1}] = (n-1)\mathbb{E}[Y_1] \qquad (69)$$

Since $Y_1, ..., Y_{n-1}$ are uniformly distributed on $(0, k)$, we know

$$f_{Y_1}(y) = \begin{cases} \frac{1}{k} \text{ if } y \in (0,k) \\ 0 \text{ otherwise} \end{cases} \tag{70}$$

We can use (70) to directly compute that

$$\mathbb{E}[Y_1] = \int_{-\infty}^{\infty} y f_{Y_1}(y) dy = \int_0^k y \frac{1}{k} dy = \frac{1}{k} \frac{y^2}{2} |_{y=0}^k = \frac{1}{k} \frac{k^2}{2} = \frac{k}{2} \tag{71}$$

Plugging (71) into (69), we find that

$$\mathbb{E}_\theta[\sum_{i=1}^{n-1} Y_j] = (n-1)\frac{k}{2} \tag{72}$$

Plugging (72) and (68) into (67) yields

$$\mathbb{E}_\theta[2\frac{1}{n}\sum_{i=1}^{n} X_i | X_{(n)} = k] = \frac{2}{n}k + \frac{2}{n}\frac{(n-1)k}{2} = \frac{2k + (n-1)k}{n} = \frac{(2+n-1)k}{n} = \frac{n+1}{n}k =: g(k) \tag{73}$$

By the definition of conditional expectation as a random variable, we can substitute $X_{(n)}$ for k in (73) to find

$$\mathbb{E}_\theta[2\frac{1}{n}\sum_{i=1}^{n} X_i | X_{(n)}] = \frac{n+1}{n} X_{(n)} \tag{74}$$

This completes the proof.
Note that this result makes intuitive sense, as

$$\mathbb{E}_\theta[2\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_\theta[X_i] = \frac{2}{n} \cdot n\frac{\theta}{2} = \theta \tag{75}$$

so the Law of Iterated Expectation guarantees that

$$\mathbb{E}[\mathbb{E}_\theta[2\frac{1}{n}\sum_{i=1}^{n} X_i | X_{(n)}]] = \mathbb{E}_\theta[2\frac{1}{n}\sum_{i=1}^{n} X_i] = \theta \tag{76}$$

and we already know from part (b) and (74) that

$$\mathbb{E}[\mathbb{E}_\theta[2\frac{1}{n}\sum_{i=1}^{n} X_i | X_{(n)}]] = \mathbb{E}[\frac{n+1}{n} X_{(n)}] = \theta \tag{77}$$

Exercise 15. Let $X_1, \ldots, X_n$ be a random sample of size n from the Bernoulli distribution with $0 < \theta < 1$ unknown. (So, $\mathbb{P}(X_1 = 1) = \theta$ and $\mathbb{P}(X_1 = 0) = 1 - \theta$.)

In class, we showed that $\sum_{i=1}^n X_i$ is consistent for θ, and also that

$$\mathbb{E}_\theta\Big(X_1 \,\Big|\, \sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

That is, the Rao-Blackwell Theorem suggests that the sample mean has small variance among all unbiased estimators for θ.

- Compute the Fisher information $I_{X_1}(\theta)$.
- Compute the Fisher information $I_{(X_1,\ldots,X_n)}(\theta)$.

- Show that $\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\theta(1-\theta)}{n}$.
- Does the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ achieve equality in the Cramer-Rao inequality? If so, then $\frac{1}{n}\sum_{i=1}^{n} X_i$ is UMVU.

Solution.
First, we will compute a few values which will prove useful in various parts of this exercise. Note that, since $X_1, ..., X_n$ are independent and identically distributed, we have

$$\mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n] \tag{78}$$

Since X_1 is a Bernoulli distributed random variable with parameter $0 < \theta < 1$, we know $\mathbb{P}(X_1 \in \{0, 1\}) = 1$, so we have

$$\mathbb{E}[X_1] = \sum_{x=0}^{1} x\mathbb{P}(X_1 = x) = 0 \cdot \mathbb{P}(X_1 = 0) + 1 \cdot \mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = 1) = \theta \tag{79}$$

and

$$\mathbb{E}[X_1^2] = \sum_{x=0}^{1} x^2\mathbb{P}(X_1 = x) = 0^2 \cdot \mathbb{P}(X_1 = 0) + 1^2 \cdot \mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = 1) = \theta \tag{80}$$

Now let $Y := \sum_{i=1}^{n} X_i = X_1 + \cdots + X_n$. Then Y is binomial random variable with parameters n and θ. Applying Linearity of Expectation and (78), we have

$$\mathbb{E}[Y] = \mathbb{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n\mathbb{E}[X_1] = n\theta \tag{81}$$

Also, since $Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ for all random variables X, we know

$$\mathbb{E}[Y^2] = Var(Y) + \mathbb{E}[Y]^2 \tag{82}$$

Using the fact that $Var(A+B) = Var(A) + Var(B)$ for all independent random variables A and B, we can directly compute that

$$Var(Y) = Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) = nVar(X_1) = n(\mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2) = n(\theta - \theta^2) = n\theta(1-\theta) \tag{83}$$

Plugging the results from (83) and (81) into (82) yields

$$\mathbb{E}[Y^2] = n\theta(1-\theta) + n^2\theta^2 \tag{84}$$

Now, we can proceed with the individual sections of the exercise.

(a) By definition of the Fisher Information, we have

$$I_{X_1}(\theta) := \mathbb{E}_\theta[(\frac{d}{d\theta} ln f_\theta(X_1))^2] \tag{85}$$

Since X_1 is a Bernoulli distributed random variable with parameter $0 < \theta < 1$, we know X_1 has PMF

$$f_\theta(X_1) = \begin{cases} \theta \text{ if } X_1 = 1 \\ 1 - \theta \text{ if } X_1 = 0 \\ 0 \text{ otherwise} \end{cases} = \theta^{X_1}(1-\theta)^{1-X_1} 1_{X_1 \in \{0,1\}}$$

so we can compute that

$$\begin{aligned}\frac{d}{d\theta}(lnf_\theta(X_1)) = &\quad 1_{X_1\in\{0,1\}}\frac{d}{d\theta}ln(\theta^{X_1}(1-\theta)^{1-X_1}) = 1_{X_1\in\{0,1\}}\frac{d}{d\theta}(X_1ln(\theta)+(1-X_1)ln(1-\theta))\\ = &\quad 1_{X_1\in\{0,1\}}(\frac{X_1}{\theta}-\frac{1-X_1}{1-\theta}) \qquad (86)\end{aligned}$$

Since $\mathbb{P}(X_1 \in \{0,1\}) = 1$ by definition of a Bernoulli distributed random variable, we know $\mathbb{P}(1_{X_1\in\{0,1\}} = 1) = 1$. That is, there is a 100% probability that $1_{X_1\in\{0,1\}} = 1$, so with 100% probability, we can rewrite (86) as

$$\frac{d}{d\theta}(lnf_\theta(X_1)) = 1(\frac{X_1}{\theta}-\frac{1-X_1}{1-\theta}) = \frac{X_1}{\theta}-\frac{1-X_1}{1-\theta} \qquad (87)$$

Squaring (87) yields

$$\begin{aligned}(\frac{d}{d\theta}(lnf_\theta(X_1)))^2 = &\qquad (\frac{X_1}{\theta}-\frac{1-X_1}{1-\theta})^2 = \frac{X_1^2}{\theta^2}-2\frac{X_1}{\theta}\frac{1-X_1}{1-\theta}+\frac{(1-X_1)^2}{(1-\theta)^2}\\ = &\qquad \frac{X_1^2}{\theta^2}-2\frac{X_1}{\theta}\frac{1-X_1}{1-\theta}+\frac{1-2X_1+X_1^2}{(1-\theta)^2}\\ = &\qquad \frac{X_1^2}{\theta^2}-2\frac{X_1-X_1^2}{\theta(1-\theta)}+\frac{1-2X_1+X_1^2}{(1-\theta)^2} \qquad (88)\end{aligned}$$

Taking the expectation of (88) and applying Linearity of Expectation yields

$$\begin{aligned}\mathbb{E}_\theta[(\frac{d}{d\theta}(lnf_\theta(X_1)))^2] = &\quad \mathbb{E}_\theta[\frac{X_1^2}{\theta^2}-2\frac{X_1-X_1^2}{\theta(1-\theta)}+\frac{1-2X_1+X_1^2}{(1-\theta)^2}]\\ = &\quad \frac{1}{\theta^2}\mathbb{E}[X_1^2]-\frac{2}{\theta(1-\theta)}(\mathbb{E}[X_1]-\mathbb{E}[X_1^2])+\frac{1}{(1-\theta)^2}(1-2\mathbb{E}[X_1]+\mathbb{E}[X_1^2]) \qquad (89)\end{aligned}$$

Plugging the results from (79) and (80) into (89) yields

$$\begin{aligned}\mathbb{E}_\theta[(\frac{d}{d\theta}(lnf_\theta(X_1)))^2] = &\quad \frac{1}{\theta^2}\theta-\frac{2}{\theta(1-\theta)}(\theta-\theta)+\frac{1}{(1-\theta)^2}(1-2\theta+\theta)\\ = &\quad \frac{1}{\theta}+0+\frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta}+\frac{1}{1-\theta} = \frac{(1-\theta)}{\theta(1-\theta)}+\frac{\theta}{\theta(1-\theta)} = \frac{1}{\theta(1-\theta)} \qquad (90)\end{aligned}$$

Comparing (90) with (85), we find that

$$I_{X_1}(\theta) := \mathbb{E}_\theta[(\frac{d}{d\theta}(lnf_\theta(X_1)))^2] = \frac{1}{\theta(1-\theta)} \qquad (91)$$

which completes part (a).

(b) By the definition of the Fisher Information, we have

$$I_{(X_1,...,X_n)}(\theta) := \mathbb{E}_\theta[(\frac{d}{d\theta}(lnf_\theta(X_1,...,X_n)))^2] \qquad (92)$$

Since $X_1, ..., X_n$ are independent, we have

$$f_\theta(X_1,...,X_n) = \prod_{i=1}^{n} f_\theta(X_i) \qquad (93)$$

Since $X_1, ..., X_n$ are identically distributed Bernoulli random variables with parameter $0 < \theta < 1$, we know

$$f_\theta(X_i) = \begin{cases}\theta \text{ if } X_i = 1\\ 1-\theta \text{ if } X_i = 0 \quad = \theta_i^X(1-\theta)^{1-X_i}1_{X_i\in\{0,1\}}\\ 0 \text{ otherwise.}\end{cases} \qquad (94)$$

Plugging the closed form expression from (94) into (93) yields

$$f_\theta(X_1,...,X_n) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} 1_{X_i \in \{0,1\}} = 1_{X_1,...,X_n \in \{0,1\}} \theta^{X_1+\cdots+X_n}(1-\theta)^{n-(X_1+\cdots+X_n)} \tag{95}$$

Since $X_1,...,X_n$ are *i.i.d.* Bernoulli distributed random variable, we know $\mathbb{P}(X_1,...,X_n \in \{0,1\}) = 1$, so we know $\mathbb{P}(1_{X_1,...,X_n \in \{0,1\}} = 1) = 1$. Thus, with 100% probability, we have

$$f_\theta(X_1,...,X_n) = \theta^{X_1+\cdots+X_n}(1-\theta)^{n-(X_1+\cdots+X_n)} \tag{96}$$

Taking the natural log of (96) yields

$$\begin{aligned} ln(f_\theta(X_1,...,X_n)) = &\quad ln(\theta^{X_1+\cdots+X_n}(1-\theta)^{n-(X_1+\cdots+X_n)}) \\ = &\quad (X_1+\cdots+X_n)ln(\theta) + (n-(X_1+\cdots+X_n))ln(1-\theta) \end{aligned} \tag{97}$$

Differentiating (97) with respect to θ yields

$$\begin{aligned} \frac{d}{d\theta}(ln(f_\theta(X_1,...,X_n))) = &\quad \frac{d}{d\theta}((X_1+\cdots+X_n)ln(\theta) + (n-(X_1+\cdots+X_n))ln(1-\theta)) \\ = &\quad \frac{X_1+\cdots+X_n}{\theta} - \frac{n-(X_1+\cdots+X_n)}{(1-\theta)} \end{aligned} \tag{98}$$

To simplify (98), we can use our definition of $Y := \sum_{i=1}^{n} X_i = X_1+\cdots+X_n$ to find

$$\frac{d}{d\theta}(ln(f_\theta(X_1,...,X_n))) = \frac{Y}{\theta} - \frac{n-Y}{1-\theta} \tag{99}$$

Squaring (99) yields

$$\begin{aligned} (\frac{d}{d\theta}(ln(f_\theta(X_1,...,X_n))))^2 = &\quad (\frac{Y}{\theta} - \frac{n-Y}{1-\theta})^2 \\ = &\quad \frac{Y^2}{\theta^2} - 2\frac{Y(n-Y)}{\theta(1-\theta)} + \frac{(n-Y)^2}{(1-\theta^2} \\ = &\quad \frac{Y^2}{\theta^2} - 2\frac{nY-Y^2}{\theta(1-\theta)} + \frac{(n-Y)^2}{(1-\theta)^2} \end{aligned} \tag{100}$$

Taking the expectation of (100) and applying Linearity of Expectation yields

$$\begin{aligned} \mathbb{E}[(\frac{d}{d\theta}(ln(f_\theta(X_1,...,X_n))))^2] = &\quad \mathbb{E}[\frac{Y^2}{\theta^2} - 2\frac{nY-Y^2}{\theta(1-\theta)} + \frac{(n-Y)^2}{(1-\theta)^2}] \\ = &\quad \frac{1}{\theta^2}\mathbb{E}[Y^2] - \frac{2}{\theta(1-\theta)}(n\mathbb{E}[Y] - \mathbb{E}[Y^2]) + \frac{1}{(1-\theta)^2}\mathbb{E}[(n-Y)^2] \end{aligned} \tag{101}$$

Since Y is a Binomial distributed random variable with parameters n and $0 < \theta < 1$, $\mathbb{P}(Y \in \{0,1,...,n\}) = 1$, so Y has PMF

$$f_Y(k) = \mathbb{P}(Y=k) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{102}$$

We can use (102) to note that

$$\mathbb{P}(n-Y=k) = \mathbb{P}(Y=n-k) = \binom{n}{n-k}\theta^{n-k}(1-\theta)^k = \binom{n}{k}(1-\theta)^k\theta^{n-k} = \mathbb{P}(Z=k) \tag{103}$$

where Z is a Binomial distributed random variable with parameters n and $0 < 1-\theta < 1$. That is, $Z = n - Y$ is the sum of n *i.i.d.* Bernoulli random variables $Z_1, ..., Z_n$ with parameter $1-\theta$. Thus, we can compute

$$\mathbb{E}[n-Y] = \mathbb{E}[Z] = \mathbb{E}[\sum_{i=1}^{n} Z_i] = \sum_{i=1}^{n} \mathbb{E}[Z_i] = n\mathbb{E}[Z_1] = n(1-\theta) \tag{104}$$

and

$$Var(n-Y) = Var(Z) = Var(\sum_{i=1}^{n} Z_i) = \sum_{i=1}^{n} Var(X_i) = nVar(Z_1) = n(1-\theta)\theta \tag{105}$$

Combining the results of (104) and (105) yields

$$\mathbb{E}[(n-Y)^2] = Var(n-Y) + \mathbb{E}[n-Y]^2 = n\theta(1-\theta) + n^2(1-\theta)^2 \tag{106}$$

Plugging the results from (106), (84), and (81) into (101) yields

$$\begin{aligned}
& \mathbb{E}[(\frac{d}{d\theta}(ln(f_\theta(X_1, ..., X_n))))^2] \\
= \quad & \frac{1}{\theta^2}(n\theta(1-\theta) + n^2\theta^2) - \frac{2}{\theta(1-\theta)}(n^2\theta - n\theta(1-\theta) - n^2\theta^2) + \frac{1}{(1-\theta)^2}(n\theta(1-\theta) + n^2(1-\theta)^2) \\
= \quad & \frac{n(1-\theta)}{\theta} + n^2 - \frac{2n^2}{(1-\theta)} + 2n + \frac{2n^2\theta}{(1-\theta)} + \frac{n\theta}{1-\theta} + n^2 \\
= \quad & \frac{n(1-\theta)}{\theta} + \frac{n\theta}{1-\theta} + 2n^2 - 2n^2\frac{1-\theta}{1-\theta} + 2n \\
= \quad & \frac{n(1-\theta)}{\theta} + \frac{n\theta}{1-\theta} + 2n^2 - 2n^2 + 2n \\
= \quad & \frac{n(1-\theta)^2 + 2n\theta(1-\theta) + n\theta^2}{(1-\theta)(\theta)} \\
= \quad & n\frac{\theta^2 + 2\theta(1-\theta) + (1-\theta)^2}{(1-\theta)\theta} \\
= \quad & n\frac{(\theta + (1-\theta))^2}{\theta(1-\theta)} \\
= \quad & n\frac{1^2}{\theta(1-\theta)} = \frac{n}{\theta(1-\theta)}
\end{aligned} \tag{107}$$

Comparing (107) with (92), we find

$$I_{(X_1,...,X_n)}(\theta) := \mathbb{E}_\theta[(\frac{d}{d\theta}(lnf_\theta(X_1, ..., X_n)))^2] = \frac{n}{\theta(1-\theta)} \tag{108}$$

which completes part (b).

We could also note that, since $X_1, ..., X_n$ are independent, Proposition 4.26 guarantees that

$$I_{X_1,...,X_n}(\theta) = \sum_{i=1}^{n} I_{X_i}(\theta) \tag{109}$$

Since $X_1, ..., X_n$ are identically distributed, we have $I_{X_1}(\theta) = \cdots = I_{X_n}(\theta)$, so we find

$$I_{X_1,...,X_n}(\theta) = nI_{X_1}(\theta) \tag{110}$$

Plugging the result from (91) into (110) yields

$$I_{X_1,...,X_n}(\theta) = n\frac{1}{\theta(1-\theta)} = \frac{n}{\theta(1-\theta)} \qquad (111)$$

Note that the result from (111) agrees with the result from (108), verifying our solution to part (b).

(c) We will prove that

$$Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{\theta(1-\theta)}{n} \qquad (112)$$

in two ways. First, note that $\frac{1}{n}Y := \frac{1}{n}\sum_{i=1}^{n} X_i$, so we know

$$Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}Var(Y) \qquad (113)$$

Plugging the result from (83) into (113) yields

$$Var(\frac{1}{n}\sum_{i=1^n} X_i) = \frac{1}{n^2}n\theta(1-\theta) = \frac{\theta(1-\theta)}{n} \qquad (114)$$

which completes the first proof of (112). For the second proof, we directly compute that

$$Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) \qquad (115)$$

By the independence of $X_1, ..., X_n$, we have

$$Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \qquad (116)$$

By the identical distribution of $X_1, ..., X_n$, we have

$$Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}nVar(X_1) \qquad (117)$$

Simplifying and lugging the results from (79) and (80) into (117) to compute $Var(X_1)$ yields

$$Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}(\theta - \theta^2) = \frac{\theta(1-\theta)}{n} \qquad (118)$$

which completes the second proof of (112).

(d) Note that, since $X_1, ..., X_n$ are *i.i.d.* with $\mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n] = \theta$, we have

$$\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\mathbb{E}[\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}n\mathbb{E}[X_1] = \mathbb{E}[X_1] = \theta \qquad (119)$$

That is, $\frac{1}{n}\sum_{i=1}^{n} X_i$ is unbiased for θ. By the Cramer-Rao inequality, for any statstic $Z = t(X_1, ..., X_n)$ such that $\mathbb{E}[Z] = \theta$, we have

$$Var_\theta(Z) \geq \frac{1}{I_{(X_1,...,X_n)}(\theta)} \qquad \forall\theta \in \Theta \qquad (120)$$

Plugging the result from (111) into (120) yields

$$Var_\theta(Z) \geq \frac{1}{\frac{n}{\theta(1-\theta)}} = \frac{\theta(1-\theta)}{n} \qquad \forall\theta \in \Theta \qquad (121)$$

Comparing (118)/(114) with (121), we have

$$Var_\theta(Z) \geq \frac{\theta(1-\theta)}{n} = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) \qquad \forall\theta \in \Theta \qquad (122)$$

for all $Z = t(X_1, ..., X_n)$ such that $\mathbb{E}_\theta[Z] = \theta$. That is, the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ *does* achieve equality in the Cramer-Rao inequality. Since $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \theta$, we know the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ is UMVU for θ.
Note, while the comparison of (118) and (121) directly shows the equality of the Cramer-Rao inequality, we can also use the fact that equality occurs $\iff$ $\frac{d}{d\theta}ln(f_\theta(X_1, ..., X_n))$ and $\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}_\theta[\frac{1}{n}\sum_{i=1}^{n} X_i]$ are multiples of each other. From (98), we know that

$$\frac{d}{d\theta}ln(f_\theta(X_1, ..., X_n)) = \frac{X_1 + \cdots + X_n}{\theta} - \frac{n - (X_1 + \cdots + X_n)}{1-\theta} \qquad (123)$$

Distributing the coefficient to the rightmost term in the RHS of (123) yields

$$\frac{d}{d\theta}ln(f_\theta(X_1, ..., X_n)) = \frac{X_1 + \cdots + X_n}{\theta} + \frac{X_1 + \cdots + X_n}{1-\theta} - \frac{n}{1-\theta} \qquad (124)$$

Multiplying and dividing the RHS of (124) by $\frac{1}{n}$ and simplifying yields

$$\begin{aligned}
\frac{d}{d\theta}ln(f_\theta(X_1, ..., X_n)) = &\quad \frac{1}{\theta}\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{1-\theta}\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{1-\theta} \\
= &\quad (\frac{1}{\theta} + \frac{1}{1-\theta})\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{1-\theta} \\
= &\quad \frac{1}{\theta(1-\theta)}\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{\theta}{\theta(1-\theta)} = \frac{1}{\theta(1-\theta)}\big(\frac{1}{n}\sum_{i=1}^{n} X_i - \theta\big) \qquad (125)
\end{aligned}$$

Recalling (119) and comparing to (125) yields

$$\frac{d}{d\theta}ln(f_\theta(X_1, ..., X_n)) = \frac{1}{\theta(1-\theta)}(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i]) \qquad (126)$$

From (126), we see that $\frac{d}{d\theta}ln(f_\theta(X_1, ..., X_n))$ and $\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}_\theta[\frac{1}{n}\sum_{i=1}^{n} X_i]$ indeed *are* multiples of each other, so we know $\frac{1}{n}\sum_{i=1}^{n} X_i$ achieves equality in the Cramer-Rao inequality. Once again, since $\frac{1}{n}\sum_{i=1}^{n} X_i$ is unbiased for θ, this implies $\frac{1}{n}\sum_{i=1}^{n} X_i$ is UMVU for θ, which completes part (d).

Assignment 5

Mathematical Statistics 408 Steven Heilman

Please provide complete and well-written solutions to the following exercises.
Due October 26, 12PM noon PST, to be uploaded as a single PDF document to Gradescope.

Homework 5 - Emerson Kahle

Exercise 16. I believe that the number of home runs hit by an MLB baseball player in a single season satisfies a Poisson distribution with some unknown parameter $\lambda > 0$. In this exercise, let's try to find the parameter $\lambda > 0$ that best fits the data, using whatever estimation method you want (e.g. MLE is fine).

Here the data can be found from:

http://seanlahman.com/download-baseball-database/

I recommend using the 2020 Version, comma delimited version. The data is in a zip file, and home run data can be found in `Core` then `batting.csv` then the column HR.

After fitting the Poisson distribution to the data, compute the total variation distance of the data from the fitted Poisson distribution. If P, Q are two probability laws on e.g. the positive integers, then the total variation distance between P and Q is

$$||P - Q||_{\text{TV}} := \frac{1}{2} \sum_{k=0}^{\infty} |P(k) - Q(k)| .$$

Here P would be the fitted Poisson distribution, and Q would be the probability distribution corresponding to the data. If $||P - Q||_{\text{TV}}$ is close to 0, then the Poisson distribution that you found fits well to the data. If $||P - Q||_{\text{TV}}$ is far from 0 (perhaps close to 1), then the Poisson distribution that you found does not fit the data well.

Try to answer the same question as above for the number of made 3 point shots among 2020 WNBA players. Data can be found here:

(this link).

The data we are particularly interested in is the column `3P`.

Solution.
First, we will find the parameter λ that best fits the data, assuming the data satisfies a Poisson distribution. We will use maximum likelihood estimation to find such a λ.
Claim: For any *i.i.d.* random variables $X, X_1, ..., X_n$ from a Poisson distribution with parameter $\lambda > 0$, a maximum likelihood estimator for λ is

$$Y_n := \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}_n \qquad (1)$$

Proof: Note that the PDF of any $X \sim Poisson(\lambda)$ has PDF

$$f_\lambda(x) = \begin{cases} e^{-\lambda} \frac{\lambda^x}{x!} \text{ if } x \in \mathbb{Z}, x \geq 0 \\ 0 \text{ otherwise} \end{cases} \qquad (2)$$

By the definition of the likelihood function (and since $X_1, ..., X_n \geq 0, \in \mathbb{Z}$ with probability 1), we have

$$L(\lambda) := \prod_{i=1}^{n} f_\lambda(X_i) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{X_i}}{X_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n X_i}}{X_1! \cdots X_n!} \qquad (3)$$

Taking the natural logarithm of both sides of (3) yields the log-likelihood function:

$$\begin{aligned} ln(L(\lambda)) = & \quad ln\big(e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n X_i}}{X_1! \cdots X_n!}\big) = ln(e^{-n\lambda}) + ln(\lambda^{\sum_{i=1}^n X_i}) - ln(X_1! \cdots X_n!) \\ = & \quad -n\lambda + ln(\lambda) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} ln(X_i!) \qquad (4) \end{aligned}$$

Since maximizing the likelihood function is equivalent to maximizing the log-likelihood function, to find the MLE for λ, it suffices to find the value of λ that maximizes $ln(L(\lambda))$. Note that differentiating (4) with respect to λ yields

$$\frac{d}{d\lambda} ln(L(\lambda)) = \frac{\sum_{i=1}^{n} X_i}{\lambda} - n \qquad (5)$$

Setting (5) equal to 0 and solving for λ yields

$$\frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0 \implies \lambda = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad (6)$$

Thus, the log-likelihood function $ln(L(\lambda))$ as computed in (4) achieves a critical point at $\lambda = \frac{1}{n} \sum_{i=1}^{n} x_i$. Differentiating (5) once more with respect to λ yields

$$\frac{d^2}{d\lambda^2} ln(L(\lambda)) = -\frac{\sum_{i=1}^{n} X_i}{\lambda^2} \qquad (7)$$

Since $X_1, ..., X_n \geq 0$ with probability 1 and we are given $\lambda > 0$ for all Poisson random variables, we know $\frac{\sum_{i=1}^{n} X_i}{\lambda^2} \geq 0$ with probability 1, so we know

$$\frac{d^2}{d\lambda^2} ln(L(\lambda)) \leq 0 \qquad (8)$$

with probability 1. Thus, since the log-likelihood function is non-decreasing until $\lambda = \frac{1}{n} \sum_{i=1}^{n} X_i$ and non-increasing after $\lambda = \frac{1}{n} \sum_{i=1}^{n} X_i$, we know $\lambda = \frac{1}{n} \sum_{i=1}^{n} X_i$ is a global maximum of the log-likelihood function $ln(L(\lambda))$. Since maximizing the log-likelihood function $ln(L(\lambda))$ is equivalent to maximizing the likelihood function λ, this completes the proof that

$$Y_n := \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}_n$$

is a maximum likelihood estimator for λ for any random variables $X, X_1, ..., X_n$ which are independent and identically Poisson distributed with parameter $\lambda > 0$.

Thus, to estimate the parameter λ that best fits the data, using maximum likelihood estimation, we just need to calculate the mean of the Home Run data. We do so with the following MATLAB code:

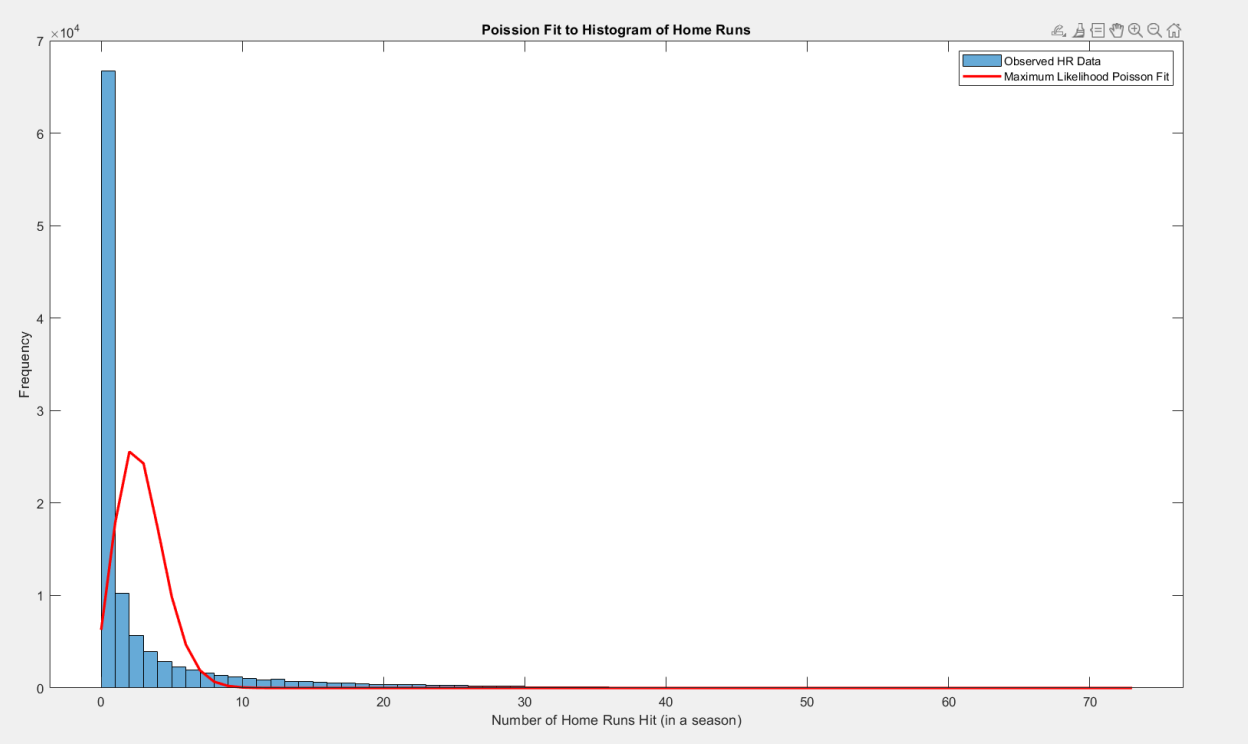
```
data = Batting.HR;
mle = mean(data);
```

Printing the variable `mle` yields our maximum likelihood estimator for the λ that best fits the Home Run data, which was `mle`$= 2.8501$ in this case. Thus, the value of the parameter λ that best fits the Home Run data is $\lambda = 2.8501$ (using maximum likelihood estimation).

We plot the PDF of a Poisson distributed random variable with parameter $\lambda = 2.8501$ on top of a histogram of the Home Run data to get a general sense of the fit of our estimated distribution. We use the following MATLAB code:

```
hist = histogram(data, 'BinWidth', 1, 'FaceColor','auto');
x=0:max(data);
poisson_pdf = poisspdf(x, mle);
hold on;
plot(x, poisson_pdf*numel(data), 'r', 'LineWidth',2);
title('Poisson Fit to Histogram of Home Runs');
xlabel('Number of Home Runs Hit (in a season)');
ylabel('Frequency');
legend('Observed HR Data', 'Maximum Likelihood Poisson Fit');
```

and observe the following output plot:



Visually, this plot suggests that our estimated Poisson distribution with parameter $\lambda = 2.8501$ does not fit the Home Run data very well.

Now, we compute $||P - Q||_{\text{TV}}$ to quantify how well the Home Run data fits the estimated Poisson distribution with parameter $\lambda = 2.8501$. To compute $||P - Q||_{\text{TV}}$, we just need to sum the absolute differences between $\mathbb{P}(X = k)$ for an $X \sim Poisson(\lambda)$ and the probability that a randomly selected player hit exactly k home runs for all $k \in \mathbb{Z}$ from 0 to ∞, then divide by two. Note that the random selection of the player implies that each player has an equal probability of being selected. Thus, the probability that a randomly selected player hit exactly k home runs is

$$\frac{|\{ \text{ players in the HR data set who hit exactly } k \text{ home runs } \}|}{|\{ \text{ players in the HR data set } \}|}$$

This expression, combined with the PDF of a Poisson distributed random variable with parameter $\lambda = 2.8501$, allows us to compute $||P - Q||_{\text{TV}}$ with the following MATLAB code:

```
tv = 0;
max1 = max(data);
freqs = zeros(max(data)+1,1);
for i=1:numel(data)
    freqs(data(i)+1) = freqs(data(i)+1) + 1;
end
for i=1:numel(freqs)
    tv = tv + abs(poissonProb(mle, i-1) - (freqs(i)/numel(data)));
end
tv = tv/2;
```

Printing `tv` yields `tv`= 0.6708, so our computed value of $||P - Q||_{\text{TV}}$ is

$$||P - Q||_{\text{TV}} \approx 0.6708$$

Since 0.6708 is much closer to 1 than it is to 0, this result suggests that our estimated Poisson distribution with parameter $\lambda = 2.8501$ does *not* fit the Home Run data very well.

We follow a very similar process for the WNBA Three Point shot data. We already showed that the maximum likelihood estimator for λ is $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ for any *i.i.d.*random variables $X, X_1, ..., X_n \sim Poisson(\lambda)$. Thus, we can estimate the parameter λ that best fits the Three Point data by computing the mean number of Three Point shots made of the Three Point data. We do so with the following MATLAB code:

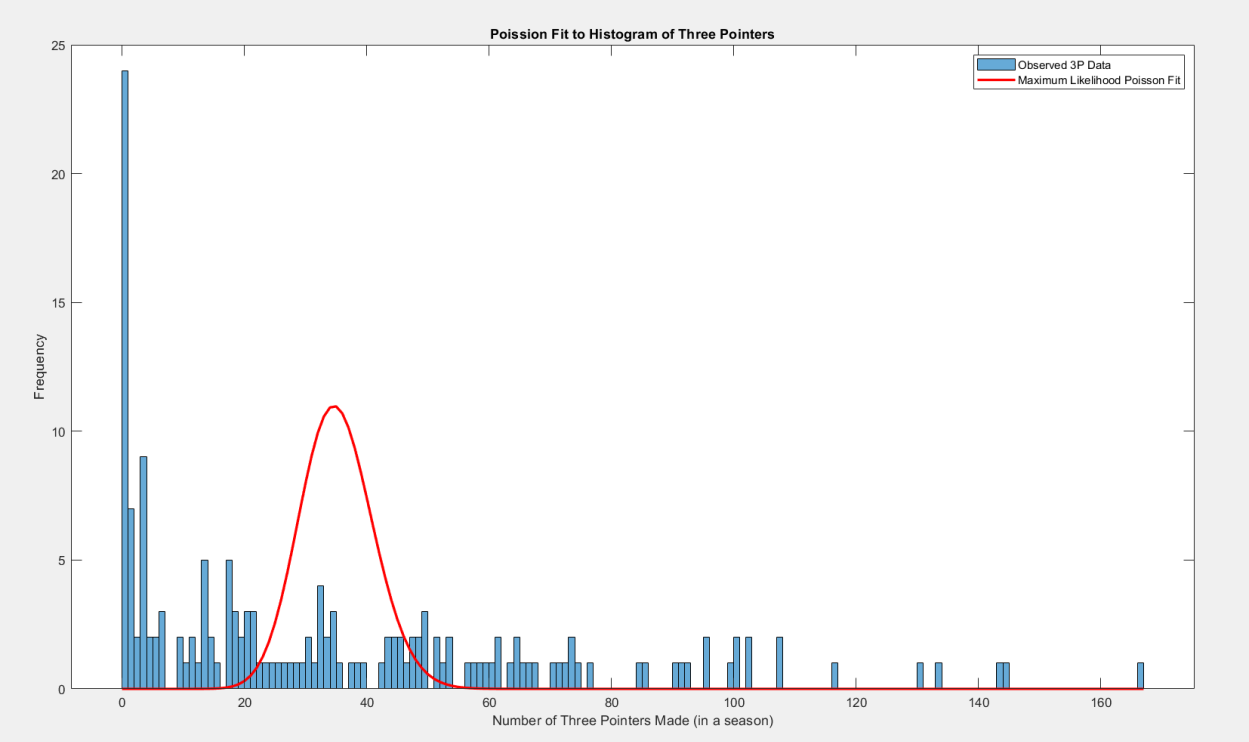
```
data2 = threePoint.PA;
mle2 = mean(data2);
```

Printing the variable `mle2` yields our maximum likelihood estimator for the λ that best fits the Three Point data, which was `mle2`= 35.1227 in this case. Thus, the value of the parameter λ that best fits the Three Point data is $\lambda = 35.1227$ (using maximum likelihood estimation).

We plot the PDF of a Poisson distributed random variable with parameter $\lambda = 35.1227$ on top of a histogram of the Three Point data to get a general sense of the fit of our estimated distribution. We use the following MATLAB code:

```
hist2 = histogram(data2, 'BinWidth', 1, 'FaceColor','auto');
x2=0:max(data2);
poisson_pdf2 = poisspdf(x2, mle2);
hold on;
plot(x2, poisson_pdf2*numel(data2), 'r', 'LineWidth',2);
title('Poisson Fit to Histogram of Three Pointers');
xlabel('Number of Three Pointers Made (in a season)');
ylabel('Frequency');
legend('Observed 3P Data', 'Maximum Likelihood Poisson Fit');
```

and observe the following plot:



Visually, this plot suggests that our estimated Poisson distribution with parameter $\lambda = 35.1227$ does *not* fit

the Three Point data very well. In fact, it appears as though this estimated Poisson distribution (with $\lambda = 35.1227$) fits the Three Point data *worse* than the previous estimated Poisson distribution (with $\lambda = 2.8501$) fit the Home Run data.

Now, we compute $||P - Q||_{\text{TV}}$ to quantify how well the Three Point data fits the estimated Poisson distribution with parameter $\lambda = 35.1227$. To compute $||P - Q||_{\text{TV}}$, we just need to sum the absolute differences between $\mathbb{P}(X = k)$ for an $X \sim Poisson(\lambda)$ and the probability that a randomly selected player made exactly k three pointers for all $k \in \mathbb{Z}$ from 0 to ∞, then divide by two. Note that the random selection of the player implies that each player has an equal probability of being selected. Thus, the probability that a randomly selected player made exactly k three pointers is

$$\frac{|\{ \text{ players in the 3P data set who made exactly } k \text{ three pointers } \}|}{|\{ \text{ players in the 3P data set } \}|}$$

This expression, combined with the PDF of a Poisson distributed random variable with parameter $\lambda = 35.1227$, allows us to compute $||P - Q||_{\text{TV}}$ with the following MATLAB code:

```
tv2 = 0;
freqs2 = zeros(max(data2)+1,1);
for i=1:numel(data2)
    freqs2(data2(i)+1) = freqs2(data2(i)+1) + 1;
end
for i=1:numel(freqs2)
    tv2 = tv2 + abs(poissonProb(mle2, i-1) - (freqs2(i)/numel(data2)));
end
tv2 = tv2/2;
```

Printing `tv2` yields `tv2`$= 0.7706$, so our computed value of $||P - Q||_{\text{TV}}$ is

$$||P - Q||_{\text{TV}} \approx 0.7706$$

Since 0.7706 is much closer to 1 than it is to 0, this result suggests that our estimated Poisson distribution with parameter $\lambda = 35.1227$ does *not* fit the Home Run data very well. Moreover, note that $0.7706 > 0.6708$, which was the value of $||P - Q||_{\text{TV}}$ for the estimated Poisson distribution with $\lambda = 2.8501$ and the Home Run data. Thus, our computed value of $||P - Q||_{\text{TV}}$ supports our visual observation that this estimated Poisson distribution (with $\lambda = 35.1227$) fits the Three Point data *worse* than the previous estimated Poisson distribution (with $\lambda = 2.8501$) fit the Home Run data.

Exercise 17. Wikipedia has a list of best selling video games with at least 10 million units sold, sorted by the number of units sold:

(this link)

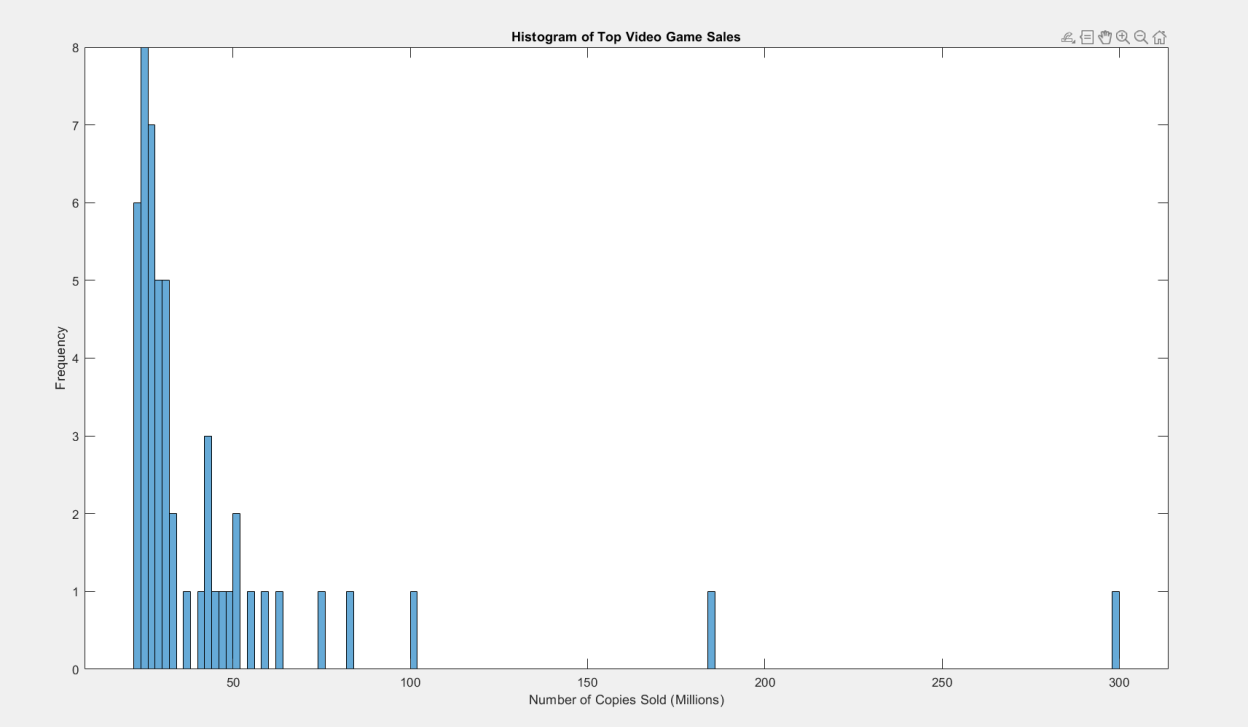
This is an open-ended question, related to this list.

Plot a histogram of the player counts of this list of games (i.e. the second column of the table). Does this histogram resemble any particular distribution? If so, try to fit that distribution to the data, as in the previous question, and use the total variation distance as a measure of goodness of fit.

Solution.
After importing the Video Game Sales data, we create the histogram with the following MATLAB code:

```
videoGames = Listofbestsellingvideogames1.Sales;
videoGames = videoGames./1e6;
histogram(videoGames, 'BinWidth', 2);
```

Note that we divide each value in the data by 1,000,000, as my computer does not have sufficient memory to compute and store the PDF of a Poisson distributed random variable for all integer values from 0 to 300,000,000. This code produces the following plot:



Note that this histogram visually resembles the PDF of a Poisson distributed random variable. Thus, we can try to fit the Video Game data to a Poisson distribution with some parameter $\lambda > 0$.

We proceed the same way as in **Exercise 1**. We already showed that the maximum likelihood estimator for λ is $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ for any $i.i.d.$random variables $X, X_1, ..., X_n \sim Poisson(\lambda)$. Thus, we can estimate the parameter λ that best fits the Video Game data by computing the mean number of Video Games sold (in millions of copies) from the Video Game data. We do so with the following MATLAB code:

```
vgMle = mean(videoGames);
```

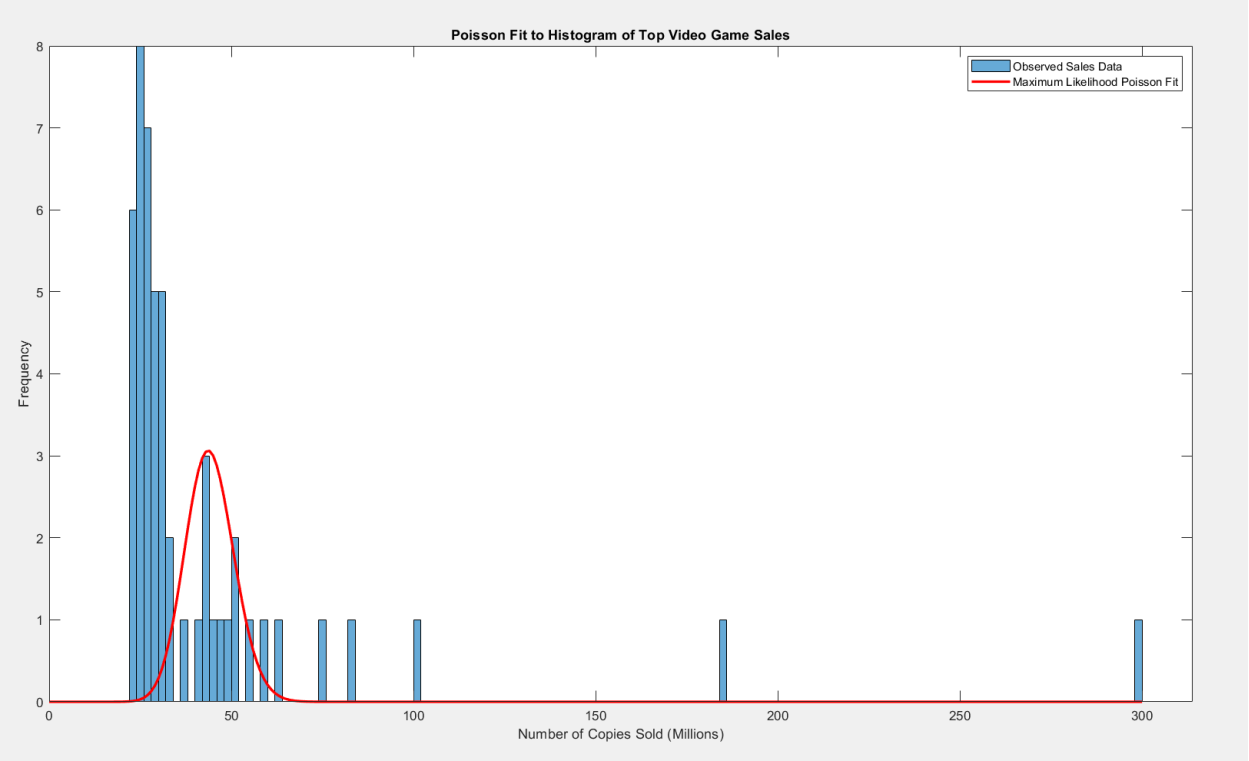
Printing the variable `vgMle` yields our maximum likelihood estimator for the λ that best fits the Video Game data, which was `vgMle`$= 44.1546$ million copies in this case. Thus, the value of the parameter λ that best fits the Video Game data is $\lambda = 44.1546$ (using maximum likelihood estimation).

We plot the PDF of a Poisson distributed random variable with parameter $\lambda = 44.1546$ on top of a histogram of the Video Game data to get a general sense of the fit of our estimated distribution. We use the following MATLAB code:

```
videoGames = Listofbestsellingvideogames1.Sales;
videoGames = videoGames./1e6;
histogram(videoGames, 'BinWidth', 2);
vgMle = mean(videoGames);
rangeVg = 0:max(videoGames);
poissonVg = poisspdf(rangeVg, vgMle);
hold on;
plot(rangeVg, poissonVg*numel(videoGames), 'r', 'LineWidth', 2);
title('Poisson Fit to Histogram of Top Video Game Sales');
xlabel('Number of Copies Sold (Millions)');
```

```
ylabel('Frequency');
legend('Observed Sales Data', 'Maximum Likelihood Poisson Fit');
```

which yields the following plot:



Visually, this plot suggests that our estimated Poisson distribution with parameter $\lambda = 44.1546$ does *not* fit the Video Game data very well.

Now, we compute $||P - Q||_{\text{TV}}$ to quantify how well the Video Game data fits the estimated Poisson distribution with parameter $\lambda = 44.1546$. To compute $||P - Q||_{\text{TV}}$, we just need to sum the absolute differences between $\mathbb{P}(X = k)$ for an $X \sim Poisson(\lambda)$ and the probability that a randomly selected video game sold exactly k million copies (when rounded to the nearest million) for all $k \in \mathbb{Z}$ from 0 to ∞, then divide by two. Note that the random selection of the video game implies that each video game has an equal probability of being selected. Thus, the probability that a randomly selected video game sold exactly k million copies, when rounded to the nearest million, is

$$\frac{|\{ \text{ video games in the VG data set who sold exactly k million copies (after rounding)}\}|}{|\{ \text{ video games in the VG data set }\}|}$$

This expression, combined with the PDF of a Poisson distributed random variable with parameter $\lambda = 44.1546$, allows us to compute $||P - Q||_{\text{TV}}$ with the following MATLAB code:

```
tvVg = 0;
videoGamesInt = int64(videoGames);
freqsVg = zeros(max(videoGames)+1);
for i=1:numel(videoGamesInt)
    freqsVg(videoGamesInt(i) + 1) = freqsVg(videoGamesInt(i) + 1) + 1;
end
for i = 1:numel(freqsVg)
    tvVg = tvVg + abs(poissonProb(vgMle, i-1) - (freqsVg(i)/numel(videoGamesInt)));
end
tvVg = tvVg/2;
```

Printing `tvVg` yields `tvVg`= 0.7310, so our computed value of $||P - Q||_{\text{TV}}$ is

$$||P - Q||_{\text{TV}} \approx 0.7310$$

Since 0.7310 is much closer to 1 than it is to 0, this result suggests that our estimated Poisson distribution with parameter $\lambda = 44.1546$ does *not* fit the Video Game data very well.

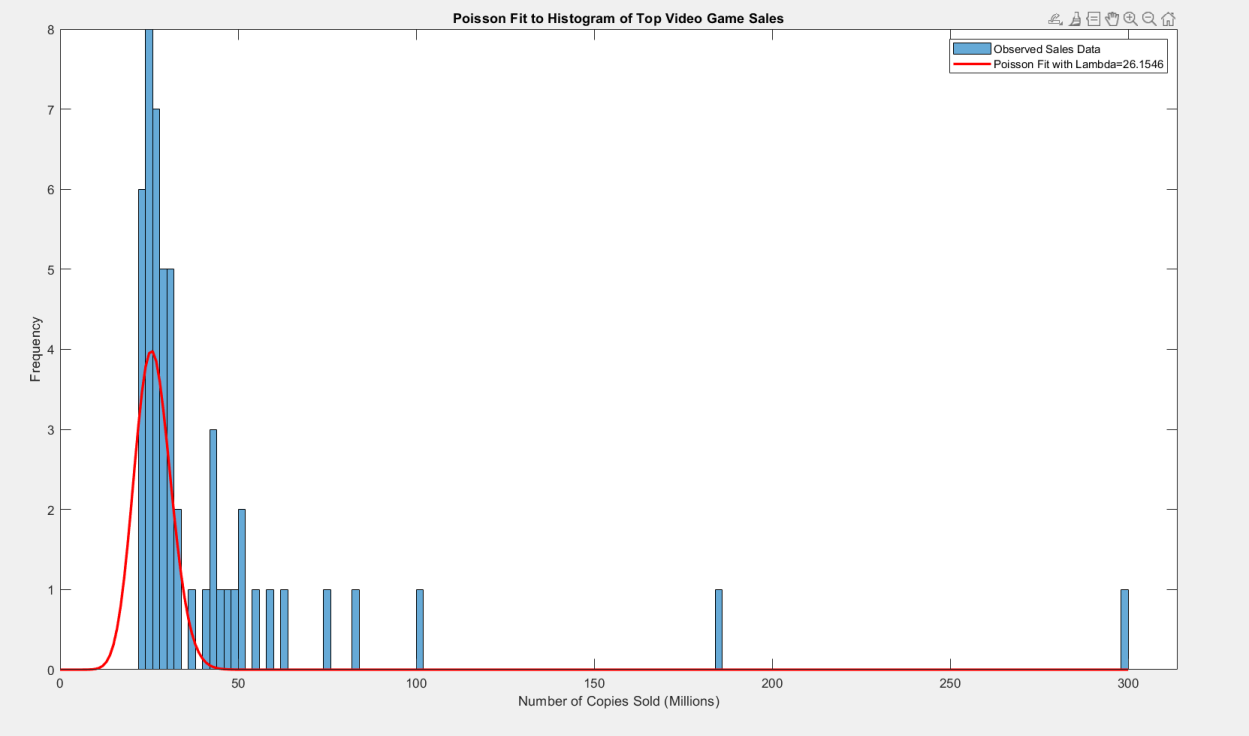
Note that the peak of the maximum likelihood Poisson distribution is approximately 18 million video games higher than the peak of the histogram of the Video Game data. Thus, intuitively, a Poisson distribution with parameter $\lambda = 44.1546 - 18 = 26.1546$ should fit the Video Game data much better than the Poisson distribution with parameter $\lambda = 44.1546$. Indeed, using the exact same MATLAB code as before, but changing the plot legends and substituting

```
vgMle = mean(videoGames) - 18;
```

for

```
vgMle = mean(videoGames)
```

we obtain the following plot:



Visually, the Poisson distribution with parameter $\lambda = 26.1546$ seems to fit the Video Game data much better. Indeed, using the same exact code (with the updated value of `vgMle`) to compute total variation distance, we find `tvVg`= 0.4066. Thus, we have

$$||P - Q||_{\text{TV}} \approx 0.4066$$

Since 0.4066 is closer to 0 than to 1, we find that the Poisson distribution with parameter $\lambda = 26.1546$ fits the Video Game data significantly better than the Poisson distribution with the maximum likelihood parameter $\lambda = 44.1546$.

In summary, while maximum likelihood estimation of the parameter λ *not* result in a Poisson distribution that fit the Video Game data very well, using a different value of λ produced a Poisson distribution which fits the Video Game data reasonably well. This completes the discussion of the fit of an estimated distribution to the Video Game data.

Exercise 18. Suppose you flip a coin 1000 times, resulting in 560 heads and 440 tails. Is it reasonable to conclude that the coin is fair (i.e. it has one half probability of heads and one half probability of tails)? Justify your answer. (Hint: if the coin is fair, then with what probability will you observe at least 560 heads being flipped? That is, is it a rare observation?)

Solution. Let $X_1, ..., X_{1000}$ such that

$$X_i = \begin{cases} 1 \text{ if the i'th coin flip is heads} \\ 0 \text{ otherwise} \end{cases} \qquad \forall i \in \{1, ..., 1000\} \qquad (9)$$

Assume that the coin is fair (i.e. $\mathbb{P}(heads) = \mathbb{P}(tails) = \frac{1}{2}$ for any randomly selected coin flip. Then $X_1, ..., X_n$ are *i.i.d.* Bernoulli distributed random variables with parameter $p = \frac{1}{2}$. That is,

$$\mathbb{P}(X_i = 1) = \frac{1}{2} = \mathbb{P}(X_i = 0) \quad \forall i \in \{1, ..., 1000\} \qquad (10)$$

Using (10), we can quickly compute that

$$\mathbb{E}[X_1] = \cdots = \mathbb{E}[X_{1000}] = \sum_{i=0}^{1} i\mathbb{P}(X_1 = i) = 1 \cdot \mathbb{P}(X_1 = 1) + 0 \cdot \mathbb{P}(X_1 = 0) = \mathbb{P}(X_1 = 1) = \frac{1}{2} \qquad (11)$$

and

$$\mathbb{E}[X_1^2] = \cdots = \mathbb{E}[X_{1000}^2] = \sum_{i=0}^{1} i^2\mathbb{P}(X_1 = i) = 1^2\mathbb{P}(X_1 = 1) + 0^2\mathbb{P}(X_1 = 0) = \mathbb{P}(X_1 = 1) = \frac{1}{2} \qquad (12)$$

We can use (11) and (12) in combination with the definition of variance to compute

$$Var(X_1) = \cdots = Var(X_{1000}) = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \frac{1}{2} - (\frac{1}{2})^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4} \qquad (13)$$

Now, recall that for all *i.i.d* random variables $X_1, ..., X_n$ with $\mathbb{E}|X_1| < \infty$ and $0 < Var(X_1) < \infty$, the Central Limit Theorem guarantees that

$$\lim_{n\to\infty} \mathbb{P}(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \qquad (14)$$

for all $-\infty < a < \infty$ where $\mu := \mathbb{E}[X_1]$ and $\sigma := \sqrt{Var(X_1)}$.
Also, note that

$$\begin{aligned} & \mathbb{P}(\text{ At least } k \text{ heads out of } n \text{ fair coin tosses }) \\ = \quad & \mathbb{P}(X_1 + \cdots + X_n \geq k) = 1 - \mathbb{P}(X_1 + \cdots + X_n \leq k - 1) \\ = \quad & 1 - \mathbb{P}(X_1 + \cdots + X_n - n\mathbb{E}[X_1] \leq k - 1 - n\mathbb{E}[X_1]) \\ = \quad & 1 - \mathbb{P}\Big(\frac{X_1 + \cdots + X_n - n\mathbb{E}[X_1]}{\sqrt{nVar(X_1)}} \leq \frac{k - 1 - n\mathbb{E}[X_1]}{\sqrt{nVar(X_1)}}\Big) \qquad (15) \end{aligned}$$

From (14), we see that, for large n,

$$P(\text{ At least } k \text{ heads out of } n \text{ fair coin tosses }) \approx \int_{-\infty}^{\frac{k-1-n\mathbb{E}[X_1]}{\sqrt{nVar(X_1)}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \qquad (16)$$

Plugging $k = 560$, $n = 1000$, $\mathbb{E}[X_1] = \frac{1}{2}$ and $Var(X_1) = \frac{1}{4}$ into (15) yields

$$\begin{aligned} & \mathbb{P}(\text{ At least 560 heads out of 1000 fair coin tosses}) \\ = \quad & 1 - \mathbb{P}\Big(\frac{X_1 + \cdots + X_n - 1000\frac{1}{2}}{\sqrt{1000\frac{1}{4}}} \leq \frac{559 - 1000\frac{1}{2}}{\sqrt{1000\frac{1}{4}}}\Big) \\ = \quad & 1 - \mathbb{P}\Big(\frac{X_1 + \cdots + X_n - 500}{5\sqrt{10}} \leq \frac{59}{5\sqrt{10}}\Big) \qquad (17) \end{aligned}$$

Since $n = 1000$ is a relatively large sample size, we know from (16) that

$$\mathbb{P}(\text{ At least 560 heads out of 1000 fair coin tosses}) \approx 1 - \int_{-\infty}^{\frac{59}{5\sqrt{10}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \tag{18}$$

Computing the right side of (18) yields

$$\mathbb{P}(\text{ At least 560 heads out of 1000 fair coin tosses}) \approx 1 - 0.999905 = 0.000095 \tag{19}$$

Thus, there is approximately a 0.0095% chance of observing at least 560 heads flipped from 1000 independent tosses of a fair coin. Therefore, it would be *very* rare to observe 560 heads out of 1000 fair coin tosses, so it is *not* reasonable to conclude that the coin is fair. If the coin was fair, we would be *very unlikely* to observe so many heads (≥ 560) from only 1000 coin tosses, so observing so many heads actually provides evidence that the coin is *not* fair. Thus, the observation of 560 heads out of 1000 coin tosses does *not* support the conclusion that the coin is fair, and it actually supports the alternative hypothesis that the coin is *not* fair (specifically, that the coin favors heads based on our calculations i.e. $\mathbb{P}(X_i = 1) > \frac{1}{2}$).

Exercise 19. Suppose the number of typos in my notes in a given year follows a Poisson distribution. In the last few years, the average number of typos was 15, and this year, I had 10 typos in my notes. Is it reasonable to conclude that the rate of typos has dropped this year? Justify your answer. (Hint: if the Poisson random variable X has a mean of 15, then with what probability will you observe that $X \leq 10$? That is, is it a rare observation?)

Solution. Following the hint, assume X is a Poisson distributed random variable with parameter $\lambda = 15 > 0$. That is,

$$\mathbb{P}(X = k) = \begin{cases} e^{-15}\frac{15^k}{k!} \text{ if } k \in \mathbb{Z}, k \geq 0 \\ 0 \text{ otherwise.} \end{cases} \tag{20}$$

so we can write

$$\mathbb{P}(X \leq 10) = \mathbb{P}(\bigcup_{i=1}^{10} X = i) = \sum_{k=0}^{10} \mathbb{P}(X = k) = \sum_{k=0}^{10} e^{-15}\frac{15^k}{k!} = e^{-15}\sum_{k=0}^{10}\frac{15^k}{k!} \tag{21}$$

Expanding (21) and directly computing yields

$$\begin{aligned} \mathbb{P}(X \leq 10) = &\quad e^{-15}(\frac{15^0}{0!} + \frac{15}{1!} + \frac{15^2}{2!} + \frac{15^3}{3!} + \frac{15^4}{4!} + \frac{15^5}{5!} + \frac{15^6}{6!} + \frac{15^7}{7!} + \frac{15^8}{8!} + \frac{15^9}{9!} + \frac{15^{10}}{10!}) \\ = &\quad e^{-15}(1 + 15 + \frac{225}{2} + \frac{15^3}{6} + \frac{15^4}{24} + \frac{15^5}{120} + \frac{15^6}{720} + \frac{15^7}{5040} + \frac{15^8}{40320} + \frac{15^9}{362880} + \frac{15^{10}}{3628800}) \\ \approx &\quad 0.118464 \end{aligned} \tag{22}$$

Thus, if the true average number of typos in the notes in a given year follows a Poisson distribution with parameter $\lambda = 15$, then there is approximately a 11.8464% chance that the number of typos in the notes in a given year will be less than or equal to 10. So, in about 1 out of every 10 years, we would expect there to be ≤ 10 typos in the notes. Thus, it is *not* reasonable to conclude that the rate of typos has dropped this year just because we observed only 10 typos this year, as observing such few typos would *not be that rare* if the true typo rate still follows a Poisson distribution with parameter $\lambda = 15$.

Alternatively, we could apply the Central Limit Theorem similarly to our application in **Exercise 3**. However, since we only have a sample of one year to observe, the Gaussian approximation of the Poisson random variable is much more crude. That being said, we proceed with this method anyway, simply to provide a sanity check for the previous work. Note that

$$\mathbb{P}(X \leq 10) = \mathbb{P}(X - \mathbb{E}[X] \leq 10 - \mathbb{E}[X]) = \mathbb{P}(\frac{X - \mathbb{E}[X]}{\sqrt{Var(X)}} \leq \frac{10 - \mathbb{E}[X]}{\sqrt{Var(X)}}) \tag{23}$$

We can use (20) to directly compute that

$$\begin{aligned}\mathbb{E}[X] = & \sum_{k=0}^{\infty} k\mathbb{P}(X=k) = \sum_{k=0}^{\infty} ke^{-15}\frac{15^k}{k!} = e^{-15}\sum_{k=1} k\frac{15^k}{k!} = e^{-15}\sum_{k=1}^{\infty}\frac{15^k}{(k-1)!} \\ = & \quad 15e^{-15}\sum_{k=1}^{\infty}\frac{15^{k-1}}{(k-1)!} = 15e^{-15}\sum_{k=0}^{\infty}\frac{15^k}{k!} = 15e^{-15}e^{15} = 15 \qquad (24)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[X^2] = & \sum_{k=0}^{\infty} k^2\mathbb{P}(X=k) = \sum_{k=0}^{\infty} k^2e^{-15}\frac{15^k}{k!} = e^{-15}\sum_{k=1}^{\infty} k^2\frac{15^k}{k!} = e^{-15}\sum_{k=1}^{\infty} k\frac{15^k}{(k-1)!} \\ = & \quad 15e^{-15}\sum_{k=1}^{\infty} k\frac{15^{k-1}}{(k-1)!} = 15e^{-15}\sum_{k=0}^{\infty}(k+1)\frac{15^k}{k!} = 15e^{-15}[\sum_{k=0}^{\infty} k\frac{15^k}{k!} + \sum_{k=0}^{\infty}\frac{15^k}{k!}] \\ = & \quad 15e^{-15}[\sum_{k=1}^{\infty} k\frac{15^k}{k!} + e^{15}] = 15e^{-15}[15\sum_{k=1}^{\infty}\frac{15^{k-1}}{(k-1)!} + e^{15}] = 15e^{-15}[15\sum_{k=0}^{\infty}\frac{15^k}{k!} + e^{15}] \\ = & \quad 15e^{-15}(15e^{15} + e^{15}) = 15^2e^{-15}e^{15} + 15e^{-15}e^{15} = 15^2 + 15 \qquad (25)\end{aligned}$$

so

$$Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 15^2 + 15 - 15^2 = 15 \qquad (26)$$

Plugging (24) and (26) into (23) and applying the Central Limit Theorem (crudely) yields

$$\mathbb{P}(X \leq 10) = \mathbb{P}(\frac{X-15}{\sqrt{15}} \leq \frac{-5}{\sqrt{15}}) \approx \int_{-\infty}^{-\frac{5}{\sqrt{15}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt \approx 0.098353 \qquad (27)$$

The result from (27) shows that the (crude) application of the Central Limit Theorem verifies that the probability of observing a single year with ≤ 10 typos in the notes is around 10%.

Regardless of which method we use to determine the probability of observing $X \leq 10$, we find that this probability is around 10%, and is thus *not that rare* if the true distribution of annual typos follows a Poisson distribution with parameter $\lambda = 15$. Thus, in both cases, the observed event is *not that rare* given the null hypothesis that $\lambda = \mathbb{E}[X] = 15$, so we do *not* have enough evidence to *reasonably* conclude that the rate of typos has dropped this year (the year in which we observed only 10 typos).

Assignment 6

Mathematical Statistics 408 Steven Heilman

Please provide complete and well-written solutions to the following exercises.

Due November 16, 12PM noon PST, to be uploaded as a single PDF document to blackboard (under the Assignments tab).

Homework 6 - Emerson Kahle

Exercise 20. Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with unknown location parameter $\theta > 0$, i.e. X_1 has density

$$g(x) := 1_{x \geq \theta} e^{-(x-\theta)}, \qquad \forall\, x \in \mathbb{R}.$$

Fix $\theta_0 \in \mathbb{R}$. Suppose we want to test that hypothesis H_0 that $\theta \leq \theta_0$ versus the alternative H_1 that $\theta > \theta_0$. That is, $\Theta = \mathbb{R}$, $\Theta_0 = \{\theta \in \mathbb{R} \colon \theta \leq \theta_0\}$ and $\Theta_0^c = \Theta_1 = \{\theta \in \mathbb{R} \colon \theta > \theta_0\}$.

- Explicitly describe the rejection region of the generalized likelihood ratio test for this hypothesis. (Hint: it might be easier to describe the region using $x_{(1)} = \min(x_1, \ldots, x_n)$.)
- (Optional) If H_0 is true, then does

$$2 \log \frac{\sup_{\theta \in \Theta} f_\theta(X_1, \ldots, X_n)}{\sup_{\theta \in \Theta_0} f_\theta(X_1, \ldots, X_n)}$$

converge in distribution to a chi-squared distribution as $n \to \infty$?

Solution.

(a) Since $X_1, ..., X_n$ are independent and identically distributed, we know their joint PDF is

$$f_\theta(x) = f_\theta(x_1, ..., x_n) = \prod_{i=1}^{n} g_\theta(x_i) = \prod_{i=1}^{n} 1_{x \geq \theta} e^{-(x-\theta)} = 1_{x_1, ..., x_n \geq \theta} e^{n\theta} e^{-(x_1 + \cdots x_n)} \tag{1}$$

for all $x = (x_1, ..., x_n) \in \mathbb{R}^n$. Define

$$x_{(1)} := \min_{i \in \{1, ..., n\}} x_i \tag{2}$$

Note that, for all $\theta \in \Theta$,

$$x_1, ..., x_n \geq \theta \iff x_{(1)} \geq \theta$$

so we can rewrite the joint PDF from (1) using the definition from (2) to find

$$f_\theta(x) = f_\theta(x_1, ..., x_n) = 1_{x_{(1)} \geq \theta} e^{n\theta} e^{-(x_1 + \cdots + x_n)} \tag{3}$$

From Definition 5.22 in the notes, the rejection region C of the generalized likelihood ratio test for this hypothesis and a constant $k \geq 1$ is

$$C := \{x \in \mathbb{R}^n : \sup_{\theta \in \Theta} f_\theta(x) \geq k \sup_{\theta in \Theta_0} f_\theta(x)\} = \{x \in \mathbb{R}^n : \frac{\sup_{\theta \in \Theta} f_\theta(x)}{\sup_{\theta \in \Theta_0} f_\theta(x)} \geq k\} \tag{4}$$

To write C explicitly, we need to individually compute the numerator and denominator from the RHS of (4). The numerator is easier to compute, so we will start there.
Fix $x = (x_1, ..., x_n) \in \mathbb{R}^n$ arbitrarily. For this arbitrary choice, consider $x_{(1)}$ as defined in (2). $f_\theta(x)$ behaves distinctly depending on whether $\theta \leq x_{(1)}$ or $\theta > x_{(1)}$. For all $\theta > x_{(1)}$, we have

$$f_\theta(x) = 1_{x_{(1)} \geq \theta} e^{n\theta} e^{-(x_1 + \cdots + x_n)} = 0 \cdot e^{n\theta} e^{-(x_1 + \cdots + x_n)} = 0 \tag{5}$$

so $f_\theta(x)$ is the constant function which always takes value 0 for all $\theta > x_{(1)}$.
On the other hand, for all $\theta \leq x_{(1)}$, we have

$$f_\theta(x) = 1_{x_{(1)} \geq \theta} e^{n\theta} e^{-(x_1 + \cdots + x_n)} = 1 \cdot e^{n\theta} e^{-(x_1 + \cdots + x_n)} = e^{n\theta} e^{-(x_1 + \cdots + x_n)} \tag{6}$$

so $f_\theta(x)$ is an exponential function in θ for all $\theta \leq x_{(1)}$. Moreover, since $e^x > 0$ for all $x \in \mathbb{R}$ (except in the limit as $x \to -\infty$), we know $e^{n\theta} > 0$ and $e^{-(x_1 + \cdots + x_n)} > 0$ for all $\theta \in \Theta$ and all $x = (x_1, ..., x_n) \in \mathbb{R}^n$. Thus, for all $\theta \leq x_{(1)}$, we have

$$f_\theta(x) = e^{n\theta} e^{-(x_1 + \cdots + x_n)} > 0 \tag{7}$$

Comparing the values of $f_\theta(x)$ in (7) and (5), we find that $f_\theta(x)$ is strictly greater for all $\theta \leq x_{(1)}$ than for all $\theta > x_{(1)}$. Thus, $f_\theta(x)$ must be maximized at some $\theta \leq x_{(1)}$. Differentiating (7), we find that, for all such $\theta \leq x_{(1)}$,

$$\frac{d}{d\theta} f_\theta(x) = n e^{n\theta} e^{-(x_1 + \cdots + x_n)} \tag{8}$$

Using the fact that n, $e^{n\theta}$, and $e^{-(x_1+\cdots+x_n)}$ are each always positive (for any $x = (x_1, ..., x_n) \in \mathbb{R}^n$ and any $\theta \in \Theta$), we can deduce from (8) that

$$\frac{d}{d\theta} f_\theta(x) > 0 \cdot 0 \cdot 0 = 0 \qquad (9)$$

for all $\theta \leq x_{(1)}$.
In summary, $f_\theta(x)$ is a strictly positive, monotonically increasing function in θ for all $\theta \leq x_{(1)}$, and $f_\theta(x)$ is a constant function taking value 0 for all $\theta > x_{(1)}$. Thus, the value that maximizes $f_\theta(x)$ for any $x = (x_1, ..., x_n) \in \mathbb{R}^n$ is the largest θ such that $f_\theta(x) \neq 0$, which is clearly $\theta = x_{(1)}$. Since $\theta > 0 \implies X_1, ..., X_n > 0$ with probability 1 ($X_i \leq 0 \implies 1_{x \geq \theta} = 0$), we know $x_{(1)} > 0$, so we know $x_{(1)} \in \Theta$. Thus, we can simplify the numerator of the RHS of (4) to be

$$\sup_{\theta \in \Theta} f_\theta(x) = f_{x_{(1)}}(x) = e^{nx_{(1)}} e^{-(x_1+\cdots+x_n)} \qquad (10)$$

Now, we will simplify the denominator of the RHS of (4). Once again, consider an arbitrary $x = (x_1, ..., x_n) \in \mathbb{R}^n$ and consider the minimum value $x_{(1)}$ as defined in (2). Then $\sup_{\theta \in \Theta_0} f_\theta(x)$ depends on whether or not $x_{(1)} \in \Theta_0$. Note that $\Theta_0 = \{\theta \in \mathbb{R} : 0 < \theta \leq \theta_0\}$, and we already know $x_{(1)} > 0$, so

$$x_{(1)} \in \Theta_0 \iff x_{(1)} \leq \theta_0 \qquad (11)$$

If $x_{(1)} \in \Theta_0$, then $f_\theta(x)$ is still a strictly positive, monotonically increasing function in θ for all $0 < \theta \leq x_{(1)}$, and $f_\theta(x)$ is still the constant function 0 for all $x_{(1)} < \theta \leq \theta_0$. Thus, for all $x \in \mathbb{R}^n$ such that $x_{(1)} \in \Theta_0$, we have

$$\sup_{\theta \in \Theta_0} f_\theta(x) = f_{x_{(1)}}(x) = e^{nx_{(1)}} e^{-(x_1+\cdots+x_n)} \qquad (12)$$

Note that $\sup_{\theta \in \Theta_0} f_\theta(x)$ is equivalent to $\sup_{\theta \in \Theta} f_\theta(x)$ for all such x. This equivalence does not hold when $x_{(1)} \notin \Theta_0$. For all such $x \in \mathbb{R}^n$ where $x_{(1)} \notin \Theta_0$, $f_\theta(x)$ is a strictly positive, monotonically increasing function in θ for all $0 < \theta \leq x_{(1)}$. Since $\theta_0 < x_{(1)}$, $f_\theta(x)$ is a strictly positive, monotonically increasing function in θ for all $0 < \theta \leq \theta_0$, or all $\theta \in \Theta_0$. Since $f_\theta(x)$ is positive and strictly increasing for all $\theta \in \Theta_0$, the $\theta \in \Theta_0$ which maximizes $f_\theta(x)$ for an arbitrary $x \in \mathbb{R}^n$ is just the maximal $\theta \in \Theta_0$. By definition, $\sup_{\theta \in \Theta_0} \theta = \theta_0$. Thus, for all $x \in \mathbb{R}^n$ such that $x_{(1)} > \theta_0$, we can simplify the denominator of the RHS of (4) to be

$$\sup_{\theta \in \Theta_0} f_\theta(x) = f_{\theta_0}(x) = e^{n\theta_0} e^{-(x_1+\cdots+x_n)} \qquad (13)$$

To achieve an explicit expression of the rejection region C, we combine the expressions for $\sup_{\theta \in \Theta_0} f_\theta(x)$ from (12) and (13) into a single expression. To do so, we make use of the following facts:

$$1-1_{x>y}, 1_{x>y} \in \{0,1\} \qquad 1-1_{x>y} = 0 \iff 1_{x>y} = 1 \qquad 1-1_{x>y} = 1 \iff 1_{x>y} = 0 \qquad (14)$$

The results from (14) allow us to express the two different cases for $\sup_{\theta \in \Theta_0} f_\theta(x)$ simultaneously using $1_{x_{(1)}>\theta}$ and $1 - 1_{x_{(1)}}$ as follows:

$$\sup_{\theta \in \Theta_0} f_\theta(x) = 1_{x_{(1)}>\theta_0} e^{n\theta_0} e^{-(x_1+\cdots+x_n)} + (1 - 1_{x_{(1)}>\theta_0}) e^{nx_{(1)}} e^{-(x_1+\cdots+x_n)} \qquad (15)$$

The results from (15) and (10) allow us to rewrite the rejection region C from (4) as

$$C = \{x \in \mathbb{R}^n : \frac{e^{nx_{(1)}} e^{-(x_1+\cdots+x_n)}}{1_{x_{(1)}>\theta_0} e^{n\theta_0} e^{-(x_1+\cdots+x_n)} + (1 - 1_{x_{(1)}>\theta_0}) e^{nx_{(1)}} e^{-(x_1+\cdots+x_n)}} \geq k\} \qquad (16)$$

for some constant $k \geq 1$. We can also express (16) as the union of the rejection region for $x \in \mathbb{R}^n$ satisfying $x_{(1)} \leq \theta_0$ and the rejection region for $x \in \mathbb{R}^n$ satisfying $x_{(1)} > \theta_0$, which simplifies the

expression. Define $A := \{x \in \mathbb{R}^n : x_{(1)} \leq \theta_0\}$ and $B := \mathbb{R}^n \setminus A = \{x \in \mathbb{R}^n : x_{(1)} > \theta_0\}$, and we can rewrite (16) as

$$\begin{aligned} C = & \quad \{x \in A : \frac{e^{nx_{(1)}}e^{-(x_1+\cdots+x_n)}}{e^{nx_{(1)}}e^{-(x_1+\cdots+x_n)}} \geq k\} \cup \{x \in B : \frac{e^{nx_{(1)}}e^{-(x_1+\cdots+x_n)}}{e^{n\theta_0}e^{-(x_1+\cdots+x_n)}} \geq k\} \\ = & \quad \{x \in A : 1 \geq k\} \cup \{x \in B : e^{n(x_{(1)}-\theta_0)} \geq k\} \\ = & \quad \{x \in A : 1 \geq k\} \cup \{x \in B : x_{(1)} - \theta_0 \geq \frac{ln(k)}{n}\} \qquad (17) \end{aligned}$$

The last line of (17) explicitly describes the rejection region of the generalized likelihood ratio test for this hypothesis for any constant $k \geq 1$. From (17), we see that, if $k = 1$, the rejection region C contains all $x \in \mathbb{R}^n$. On the other hand, if $k > 1$, the rejection region C contains all $x \in \mathbb{R}^n$ such that $x_{(1)} > \theta_0$ and $x_{(1)} - \theta_0 \geq \frac{ln(k)}{n}$. Thus, for all such $k > 1$, the rejection region includes only those $x \in \mathbb{R}^n$ whose corresponding minimum $x_{(1)}$ is sufficiently larger than θ_0.

(b) If H_0 is true,

$$2log\frac{\sup_{\theta\in\Theta} f_\theta(X_1, ..., X_n)}{\sup_{\theta\in\Theta_0} f_\theta(X_1, ..., X_n)}$$

does *not* necessarily converge in distribution to a chi-squared distribution as $n \to \infty$. From **Theorem 5.28** in the notes, we know that

$$2log\frac{\sup_{\theta\in\Theta} f_\theta(X_1, ..., X_n)}{\sup_{\theta\in\Theta_0} f_\theta(X_1, ..., X_n)}$$

converges in distribution to a chi-squared distribution as $n \to \infty$ *if* we are testing the hypothesis H_0 that $\{\theta = \theta_0\}$ versus the alternative $\{\theta \neq \theta_0\}$. In this case, however, we are testing the hypothesis H_0 that $\{\theta \leq \theta_0\}$ versus the alternative $\{\theta > \theta_0\}$. Thus, Theorem 5.28 *does not* apply, so we cannot conclude that

$$2log\frac{\sup_{\theta\in\Theta} f_\theta(X_1, ..., X_n)}{\sup_{\theta\in\Theta_0} f_\theta(X_1, ..., X_n)}$$

converges in distribution to a chi-squared distribution as $n \to \infty$.

Exercise 21. Let $X_1, \ldots, X_n$ be a random sample from a Gaussian random variable with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$.

Fix $\mu_0 \in \mathbb{R}$. Suppose we want to test that hypothesis H_0 that $\mu = \mu_0$ versus the alternative H_1 that $\mu \neq \mu_0$.

- Explicitly describe the rejection region of the generalized likelihood ratio test for this hypothesis.
- Give an explicit formula for the p-value of this hypothesis test. (Hint: If S^2 denotes the sample variance and $\overline{X}$ denotes the sample mean, you should then be able to use the statistic $\frac{(\overline{X}-\mu_0)^2}{S^2}$. Since we have an explicit formula for Snedecor's distribution, you should then be able to write an explicit integral formula for the p-value of this test.)

Solution.

(a) The PDF of a Gaussian random variable X with mean μ and variance $\sigma^2 > 0$ is

$$f_X(x) := \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \forall x \in \mathbb{R} \qquad (18)$$

Since $X_1, ..., X_n$ are independent and identically distributed, we know their joint PDF is

$$\begin{aligned} f_{\mu,\sigma^2}(x) = & \quad f_{\mu,\sigma^2}(x_1, ..., x_n) = \prod_{i=1}^n f_{X_i}(x) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ = & \quad \big(\frac{1}{\sigma\sqrt{2\pi}}\big)^n e^{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}} = \big(\frac{1}{\sigma^2 2\pi}\big)^{n/2} e^{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}} \qquad (19) \end{aligned}$$

for all $x = (x_1, ..., x_n) \in \mathbb{R}^n$. By **Definition 5.22**, the rejection region of the generalized likelihood ratio for this test is

$$C := \{x \in \mathbb{R}^n : \sup_{\theta \in \Theta} f_\theta(x) \geq k \sup_{\theta \in \Theta_0} f_\theta(x)\} = \{x \in \mathbb{R}^n : \frac{\sup_{\theta \in \Theta} f_\theta(x)}{\sup_{\theta \in \Theta_0} f_\theta(x)} \geq k\} \qquad (20)$$

In our case $\Theta = \mathbb{R} \times (0, \infty)$, as $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$. Thus, we can simplify the numerator of the RHS of (20) to be

$$\sup_{\theta \in \Theta} f_\theta(x) = \sup_{(\mu, \sigma^2) \in \Theta} \big(\frac{1}{\sigma^2 2\pi}\big)^{n/2} e^{-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}} \qquad (21)$$

The null hypothesis H_0 only imposes the restriction $\mu = \mu_0$ on Θ. Thus, $\Theta_0 = \{(\mu, \sigma^2) \in \Theta : \mu = \mu_0\} = \{(\mu_0, \sigma^2) : \sigma^2 \in (0, \infty)\}$. This allows us to simplify the denominator of the RHS of (20) to be

$$\sup_{\theta \in \Theta_0} f_\theta(x) = \sup_{(\mu_0, \sigma^2) \in \Theta_0} f_{\mu, \sigma}(x) = \sup_{\sigma^2 \in (0, \infty)} \big(\frac{1}{\sigma^2 2\pi}\big)^{n/2} e^{-\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{2\sigma^2}} \qquad (22)$$

Note that (21) is just the likelihood function evaluated at the $(\mu, \sigma^2) \in \Theta$ which maximizes it. Thus, $\sup_{\theta \in \Theta} f_\theta(x)$ is just the likelihood function evaluated at the Maximum Likelihood Estimators for μ and σ^2. Recall from **Exercise 4.41** that, for a random sample from a Gaussian distribution with unknown mean $\mu \in \mathbb{R}$ and unknown variance $\sigma^2 > 0$, the MLE for μ is

$$\hat{\mu} := \frac{1}{n} \sum_{i=1}^n X_i = \overline{X} \qquad (23)$$

and the MLE for σ^2 is

$$\hat{\sigma^2} := \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X})^2 \qquad (24)$$

Substituting the results from (23) and (24) into (21) yields

$$\begin{aligned}
\sup_{\theta \in \Theta} f_\theta(x) = & \quad \big(\frac{1}{2\pi \frac{1}{n} \sum_{i=1}^n (x_i - \overline{x})^2}\big)^{n/2} \exp\big(-\frac{\sum_{i=1}^n (x_i - \frac{1}{n} \sum_{i=1}^n x_i)^2}{2\frac{1}{n} \sum_{i=1}^n (x_i - \overline{x})^2}\big) \\
= & \quad (2\pi)^{-\frac{n}{2}} \big(\frac{1}{n} \sum_{i=1}^n (x_i - \overline{x})^2\big)^{-\frac{n}{2}} \exp\big(-\frac{\sum_{i=1}^n (x_i - \overline{X})^2}{2\frac{1}{n} \sum_{i=1}^n (x_i - \overline{x})^2}\big) \\
= & \quad (2\pi)^{-\frac{n}{2}} \big(\frac{1}{n} \sum_{i=1}^n (x_i - \overline{x})^2\big)^{-\frac{n}{2}} e^{-\frac{1}{2}\frac{}{n}} \\
= & \quad (2\pi)^{-\frac{n}{2}} \big(\frac{1}{n} \sum_{i=1}^n (x_i - \overline{x})^2\big)^{-\frac{n}{2}} e^{-\frac{n}{2}} \qquad (25)
\end{aligned}$$

Similarly, (22) is just the likelihood function evaluated at the $(\mu_0, \sigma^2) \in \Theta_0$ which maximizes it. Since $\mu = \mu_0$ is fixed, $\sup_{\theta \in \Theta_0} f_\theta(x)$ is just the likelihood function evaluated at μ_0 and the MLE for σ^2 given $\mu = \mu_0$.
Note that taking the natural log of the likelihood function from (19) with $\mu = \mu_0$ known, σ^2 unknown is

$$ln(f_{(\mu_0, \sigma^2)}(x)) = -nln(2\pi) - nln(\sigma) - \frac{\sum_{i=1}^n (x_i - \mu_0)^2}{2\sigma^2} \qquad (26)$$

Differentiating (26) with respect to σ yields

$$\frac{d}{d\sigma} ln(f_{(\mu_0, \sigma^2)}(x)) = \frac{-n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (x_i - \mu_0)^2 \qquad (27)$$

Setting (27) equal to 0 and solving for σ^2 yields

$$\frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu_0)^2 = 0 \iff \frac{n}{\sigma} = \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu_0)^2 \iff \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2 \tag{28}$$

Thus, our MLE for σ^2 with known mean $\mu = \mu_0$ is

$$\hat{\sigma^2}^* = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_0)^2 \tag{29}$$

Substituting μ_0 for μ and $\hat{\sigma^2}^*$ for σ^2 in (22) yields

$$\begin{aligned}
\sup_{\theta\in\Theta_0} f_\theta(x) = & \quad \Big(\frac{1}{2\pi\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2}\Big)^{n/2} e^{-\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{2\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2}} \\
= & \quad (2\pi)^{-\frac{n}{2}}\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2\Big)^{-\frac{n}{2}} \exp\big(-\frac{1}{\frac{2}{n}}\big) \\
= & \quad (2\pi)^{-\frac{n}{2}}\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2\Big)^{-\frac{n}{2}} e^{-\frac{n}{2}}
\end{aligned} \tag{30}$$

Plugging the results from (25) and (30) into (20) yields

$$\begin{aligned}
C = & \quad \{x\in\mathbb{R}^n : \frac{(2\pi)^{-\frac{n}{2}}\big(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\big)^{-\frac{n}{2}} e^{-\frac{n}{2}}}{(2\pi)^{-\frac{n}{2}}\big(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2\big)^{-\frac{n}{2}} e^{-\frac{n}{2}}} \geq k\} \\
= & \quad \{x\in\mathbb{R}^n : \frac{\big(\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\big)^{-\frac{n}{2}}}{\big(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2\big)^{-\frac{n}{2}}} \geq k\} \\
= & \quad \{x\in\mathbb{R}^n : \Big(\frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\Big)^{\frac{n}{2}} \geq k\} \\
= & \quad \{x\in\mathbb{R}^n : \frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \geq k^{\frac{2}{n}}\}
\end{aligned} \tag{31}$$

While the last line of (31) does provide an explicit expression for the generalized likelihood ratio test for this hypothesis with a constant $k \geq 1$, but we can simplify it further to make the expression of C more informative.
First, note that

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-\mu_0)^2 = & \quad \sum_{i=1}^{n}((x_i-\overline{x})+(\overline{x}-\mu_0))^2 = \sum_{i=1}^{n}((x_i-\overline{x})^2+(\overline{x}-\mu_0)^2+2(x_i-\overline{x})(\overline{x}-\mu_0)) \\
= & \quad \sum_{i=1}^{n}(x_i-\overline{x})^2 + \sum_{i=1}^{n}(\overline{x}-\mu_0)^2 = \big(\sum_{i=1}^{n}(x_i-\overline{x})^2\big) + n(\overline{x}-\mu_0)^2 + 2(\overline{x}-\mu_0)\sum_{i=1}^{n}\overline{x}-x_i \\
= & \quad \big(\sum_{i=1}^{n}(x_i-\overline{x})^2\big) + n(\overline{x}-\mu_0)^2 + 2(\overline{x}-\mu_0)(n\overline{x}-\sum_{i=1}^{n}x_i) \\
= & \quad \big(\sum_{i=1}^{n}(x_i-\overline{x})^2\big) + n(\overline{x}-\mu_0)^2 + 2(\overline{x}-\mu_0)(\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}x_i) \\
= & \quad \big(\sum_{i=1}^{n}(x_i-\overline{x})^2\big) + n(\overline{x}-\mu_0)^2
\end{aligned} \tag{32}$$

Now, we can simplify the fraction from the last line of (31) to find

$$\frac{\sum_{i=1}^n (x_i-\mu_0)^2}{\sum_{i=1}^n (x_i-\overline{x})^2} = \frac{\left(\sum_{i=1}^n (x_i-\overline{x})^2\right) + n(\overline{x}-\mu_0)^2}{\sum_{i=1}^n (x_i-\overline{x})^2} = 1 + \frac{n(\overline{x}-\mu_0)^2}{\sum_{i=1}^n (x_i-\overline{x})^2} \qquad (33)$$

Using (33), we can rewrite (31) as

$$\begin{aligned} C = & \quad \{x \in \mathbb{R}^n : 1 + \frac{n(\overline{x}-\mu_0)^2}{\sum_{i=1}^n (x_i-\overline{x})^2} \geq k^{\frac{2}{n}}\} \\ = & \quad \{x \in \mathbb{R}^n : \frac{n(\overline{x}-\mu_0)^2}{\sum_{i=1}^n (x_i-\overline{x})^2} \geq k^{\frac{2}{n}} - 1\} \\ = & \quad \{x \in \mathbb{R}^n : \frac{(\overline{x}-\mu_0)^2}{\sum_{i=1}^n (x_i-\overline{x})^2} \geq \frac{1}{n}(k^{\frac{2}{n}} - 1)\} \qquad (34) \end{aligned}$$

Note that the sample variance satisfies

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i-\overline{x})^2 \qquad (35)$$

so we can multiply both sides of the inequality from (34) by (n-1) to find

$$\begin{aligned} C = & \quad \{x \in \mathbb{R}^n : \frac{(\overline{x}-\mu_0)^2}{\frac{1}{n-1}\sum_{i=1}^n (x_i-\overline{x})^2} \geq \frac{n-1}{n}(k^{\frac{2}{n}} - 1)\} \\ = & \quad \{x \in \mathbb{R}^n : \frac{(\overline{x}-\mu_0)^2}{S^2} \geq \frac{n-1}{n}(k^{\frac{2}{n}} - 1)\} \qquad (36) \end{aligned}$$

The last line of (36) provides a more informative explicit expression for the generalized likelihood ratio test for this hypothesis with a constant $k \geq 1$, completing part (a). We will utilize the information provided by this expression in part (b).

(b) We will use the statistic $t : \mathbb{R}^n \to \mathbb{R}$, defined by

$$t(x) = t(x_1, ..., x_n) = \frac{(\frac{1}{n}\sum_{i=1}^n x_i - \mu_0)^2}{\frac{1}{n-1}\sum_{i=1}^n (x_i - \frac{1}{n}\sum_{i=1}^n x_i)^2} = \frac{(\overline{x}-\mu_0)^2}{S^2} \qquad (37)$$

for all $x = (x_1, ..., x_n) \in \mathbb{R}^n$. This allows us to rewrite the rejection region C from (36) as

$$C = \{x \in \mathbb{R}^n : t(x) \geq \frac{n-1}{n}(k^{\frac{2}{n}} - 1) \qquad (38)$$

Since n and k are constants, $c = \frac{n-1}{n}(k^{\frac{2}{n}} - 1)$ is a constant. Thus, by definition of the p-value, we have

$$p(x) := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(t(X) \geq t(x)) = \sup_{(\mu_0, \sigma^2) \in \Theta_0} \mathbb{P}_{(\mu_0,\sigma^2)}(\frac{(\overline{X}-\mu_0)^2}{S_X^2} \geq \frac{(\overline{x}-\mu_0)^2}{S_x^2}) \qquad (39)$$

Note that for all $\theta = (\mu, \sigma^2) \in \Theta_0$, $\mu = \mu_0$, so $X_1, ..., X_n$ are *i.i.d.* Gaussian random variables with known mean μ_0 and unknown variance $\sigma^2 > 0$. This implies

$$\mathbb{E}[X_1] = \cdots = \mathbb{E}[X_n] = \mathbb{E}[\overline{X}] = \mu_0 \qquad (40)$$

Since the sum of Gaussians random variables is also a Gaussian random variable, we know $\overline{X}$ is also a Gaussian random variable, also with mean μ_0, although it has variance

$$Var(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n} \qquad (41)$$

Thus, $\overline{X} - \mu_0$ is a Gaussian random variable with mean 0 and variance $\frac{\sigma^2}{n}$, so dividing by by $\frac{\sigma}{\sqrt{n}}$ yields a Gaussian random variable with mean μ_0 and variance

$$Var(\frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}) = \frac{n}{\sigma^2} Var(\overline{X} - \mu_0) = \frac{n}{\sigma^2}\frac{\sigma^2}{n} = 1 \qquad (42)$$

That is, $\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma}$ is a standard Gaussian random variable. By definition of the chi-squared distribution, this means

$$\chi^* := \big(\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}\big)^2 \sim \chi^2(1) \qquad (43)$$

so χ^* is a chi-squared random variable with 1 degree of freedom. Note that

$$\chi^* := \big(\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}\big)^2 = \frac{n}{\sigma^2}(\overline{X} - \mu_0)^2 \qquad (44)$$

so the numerator of our statistic $t(X)$ from (39) can be written as

$$(\overline{X} - \mu_0)^2 = \frac{\sigma^2}{n}\chi^* \qquad (45)$$

Also, from **Proposition 3.7**, we know that

$$\chi^\# := \frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2(n-1) \qquad (46)$$

so the denominator of our statistic $t(X)$ from (39) can be written as

$$S_X^2 = \frac{\sigma^2}{n-1}\chi^\# \qquad (47)$$

Combining (45) and (47), we can rewrite our entire statistic $t(X)$ from (39) as

$$t(X) = \frac{(\overline{X} - \mu_0)^2}{S_X^2} = \frac{\frac{\sigma^2}{n}\chi^*}{\frac{\sigma^2}{n-1}\chi^\#} = \frac{1}{n}\frac{\frac{\chi^*}{1}}{\frac{\chi^\#}{n-1}} \qquad (48)$$

Also, from **Proposition 3.7**, we know that $\overline{X}$ and S_X are independent, so we know χ^* and $\chi^\#$ are independent. By definition, for two independent chi-squared random variables Y and Z with p and q degrees of freedom, respectively, $\frac{\frac{Y}{p}}{\frac{Z}{q}}$ is a Snedecor's f-distributed random variable with p and q degrees of freedom. Since χ^* is a chi-squared random variable with 1 degree of freedom and $\chi^\#$ is a chi-squared random variable, we know

$$t(X) = \frac{1}{n}\frac{\frac{\chi^*}{1}}{\frac{\chi^\#}{n-1}}$$

is $\frac{1}{n}$ multiplied by a Snedecor's f-distributed random variable with 1 and $n-1$ degrees of freedom for all $(\mu_0, \sigma^2) \in \Theta_0$. That is,

$$f_{nt(X)}(t) = \frac{t^{-\frac{1}{2}}(\frac{1}{n-1})^{\frac{1}{2}}\Gamma(\frac{n}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n-1}{2})}\big(1 + \frac{t}{n-1}\big)^{-\frac{n}{2}} \qquad \forall t > 0 \qquad (49)$$

Multiplying both sides of the inequality in (39) by n yields

$$p(x) = \sup_{(\mu_0,\sigma^2)\in\Theta_0} \mathbb{P}_{(\mu_0,\sigma^2)}(\frac{\frac{\chi^*}{1}}{\frac{\chi^\#}{n-1}} \geq \frac{n(\overline{x} - \mu_0)^2}{S_x^2}) \qquad (50)$$

Since $\frac{\frac{\chi^*_1}{1}}{\frac{\chi^\#_{n-1}}{n-1}}$ is a Snedecor's f-distributed random variable with 1 and $n-1$ degrees of freedom for all $(\mu_0, \sigma^2) \in \Theta_0$, we can let $F := \frac{\frac{\chi^*_1}{1}}{\frac{\chi^\#_{n-1}}{n-1}} = nt(X)$ and F has PDF

$$f_F(t) = f_{nt(X)}(t) \qquad (51)$$

as defined in (49). This is true for all $(\mu_0, \sigma^2) \in \Theta_0$, regardless of the value of $\sigma^2 \in (0, \infty)$, so we can rewrite (50) as

$$p(x) = \sup_{(\mu_0,\sigma^2)\in\Theta_0} \mathbb{P}_{(\mu_0,\sigma^2)}(\frac{\frac{\chi^*_1}{1}}{\frac{\chi^\#_{n-1}}{n-1}} \geq \frac{n(\overline{x}-\mu_0)^2}{S_x^2}) = \mathbb{P}_{1,n-1}(F \geq \frac{n(\overline{x}-\mu_0)^2}{S_x^2}) \qquad (52)$$

where $\mathbb{P}_{1,n-1}$ refers to taking the probability law over Snedecor's f-distribution with 1 and $n-1$ degrees of freedom. Note that $\frac{n(\overline{x}-\mu_0)^2}{S_x^2}$ is a constant for a fixed $x \in \mathbb{R}^n$. By definition of the probability density function, for any continuous random variable F with PDF f_F, the probability that F is at least a constant c is

$$\mathbb{P}(F \geq c) = \int_c^\infty f_F(t)dt \qquad (53)$$

Plugging the result from (53) and (51) into (52) yields

$$p(x) = \mathbb{P}_{1,n-1}(F \geq \frac{n(\overline{x}-\mu_0)^2}{S_x^2}) = \int_{\frac{n(\overline{x}-\mu_0)^2}{S_x^2}}^\infty f_F(t)dt = \int_{\frac{n(\overline{x}-\mu_0)^2}{S_x^2}}^\infty f_{nt(X)}dt \qquad (54)$$

Plugging the definition of $f_{nt(X)}$ (the PDF of a Snedecor's f-distributed random variable with 1 and $n-1$ degrees of freedom) from (49) into (54), we find

$$p(x) = \int_{\frac{n(\overline{x}-\mu_0)^2}{S_x^2}}^\infty \frac{t^{-\frac{1}{2}}(\frac{1}{n-1})^{\frac{1}{2}}\Gamma(\frac{n}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n-1}{2})}\big(1+\frac{t}{n-1}\big)^{-\frac{n}{2}}dt \qquad (55)$$

By definition of the p-value, since (55) holds for all $x \in \mathbb{R}^n$, the p-value for this hypothesis test is defined to be the statistic

$$p(X) = \int_{\frac{n(\overline{X}-\mu_0)^2}{S_X^2}}^\infty \frac{t^{-\frac{1}{2}}(\frac{1}{n-1})^{\frac{1}{2}}\Gamma(\frac{n}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n-1}{2})}\big(1+\frac{t}{n-1}\big)^{-\frac{n}{2}}dt \qquad (56)$$

The expression on the RHS of (56) is the explicit integral formula for the p-value of this hypothesis test, which completes part (b).

Exercise 22. Write down the generalized likelihood ratio estimate for the following alpha particle data, as we did in class for a slightly different data set. The corresponding test treats individual counts of alpha particles as independent Poisson random variables, versus the alternative that the probability of a count appearing in each box of data is a sequence of nonnegative numbers that sum to one. (In doing so, you should need to compute a maximum likelihood estimate using a computer.)

m	0, 1 or 2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	≥ 17
# of Intervals	16	26	58	102	125	146	163	164	120	100	72	54	20	12	10	4

Plot the MLE for the Poisson statistic (i.e. plot the denominator of the generalized likelihood ratio test statistic $\frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)}$) as a function of λ.

Finally, compute the value s of Pearson's chi-squared statistic S, and compute the probability that $S \geq s$ (assuming H_0 holds). Does the probability $\mathbb{P}(S \geq s)$ give you confidence that the null hypothesis is true?

Solution.
First, we will write down the generalized likelihood ratio estimate for the alpha particle data. By definition, the generalized likelihood ratio estimate is

$$\frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)} \tag{57}$$

so we need to find explicit expressions for Θ and Θ_0. Note that the sum of the frequencies in the table is

$$16+26+58+102+125+146+163+164+120+100+72+54+20+12+10+4=1192 \tag{58}$$

so there are 1192 time intervals with alpha particle counts split into 16 categories (columns) in the data set. Denote the alpha particle emission count over a randomly selected interval to be k. Then we can let $p_i =$ the probability that k falls in the ith column of the table, for all $1 \leq i \leq 16$. Since the columns of the table include all possible particle emission counts (i.e. all non-negative integers), for all $\theta \in \Theta$, we must have

$$\sum_{i=1}^{16} p_i = 1 \tag{59}$$

Also, since each p_i is a probability, the axioms of probability guarantee that

$$p_1, ..., p_{16} \geq 0 \tag{60}$$

The null hypothesis H_0 asserts that the counts of alpha particles in each time interval are 1192 *i.i.d.* Poisson random variables with some unknown parameter $\lambda > 0$. Thus, if H_0 is true, then we know the probability that $k = i$ is

$$q_i(\lambda) := \mathbb{P}(k=i) = e^{-\lambda}\frac{\lambda^i}{i!} \tag{61}$$

for some unknown $\lambda > 0$. Since the first column includes all particle emission counts $k \in \{0,1,2\}$, if H_0 is true, we have

$$p_1 = q_0 + q_1 + q_2 = e^{-\lambda}(1+\lambda+\frac{\lambda^2}{2}) \tag{62}$$

For all $2 \leq i \leq 15$, each column includes only those intervals with exactly $k = i+1$ alpha particle emissions. Thus, if H_0 is true, we have

$$p_i = q_{i+1} = e^{-\lambda}\frac{\lambda^{i+1}}{(i+1)!} \tag{63}$$

for all $i \in \{2, ..., 15\}$. Finally, since the 16th column includes all intervals with particle emission counts $k \in \{17, 18, ...\}$, if H_0 is true, we have

$$p_{16} = \sum_{i=17}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!} \tag{64}$$

Since (62), (63), and (64) are both necessary and sufficient for the null hypothesis to be true, we can write the null hypothesis explicitly:

$$H_0 : \theta \in \Theta_0 = \{p_1, ..., p_{16} : p_1 = q_0+q_1+q_2, p_2 = q_3, ..., p_{15} = q_{16}, p_{16} = \sum_{i=17}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!}\} \tag{65}$$

If we do not assume H_0 is true, the axioms of probability and the structure of the data set still guarantee that (59) and (60) hold, so we can also define the parameter space Θ explicitly:

$$\Theta = \{p_1, ..., p_{16} : p_1, ..., p_{16} \geq 0, \sum_{i=1}^{16} p_i = 1\} \tag{66}$$

Thus, without assuming H_0 is true, the 1192 alpha particle emission counts represent 1192 independent rolls of the same 16 sided die, where the probability of rolling an i is p_i for all $i \in \{1, ..., 16\}$ and 0 for all other i. Define $X_1, ..., X_{16}$ such that $X_i =$ the number of alpha particle emission counts in the ith column of the table from 1192 randomly selected time intervals. For each time interval, there is independently a p_i probability that the corresponding alpha particle count k falls in the ith column of the table (and a $1 - p_i$ probability that it does not). Thus, $X_1, ..., X_n$ are Binomial random variables such that X_i has parameters $n = 1192$ and $p = p_i$. The joint distribution of $X_1, ..., X_{16}$, which describes our data set, can then be modeled as a multinomial distribution whose joint PDF satisfies

$$f_\theta(x) = f_{p_1,...,p_{16}}(x_1, ..., x_{16}) = 1192! \prod_{i=1}^{16} \frac{p_i^{x_i}}{x_i!} \tag{67}$$

for all $x = (x_1, ..., x_{16})$ such that $x_1, ..., x_{16} \geq 0$ and $\sum_{i=1}^{16} x_i = 1192$. Since (67) holds for all $\theta \in \Theta$, it holds for all $\theta \in \Theta_0 \subseteq \Theta$. Thus, if H_0 is true, we have

$$\begin{aligned} f_\theta(x) = \quad & f_{p_1,...,p_{16}}(x_1, ..., x_{16}) = 1192! \prod_{i=1}^{16} \frac{p_i^{x_i}}{x_i!} \\ = \quad & 1192! \frac{(q_0 + q_1 + q_2)^{x_1}}{x_1!} \left(\prod_{i=2}^{15} \frac{q_{i+1}^{x_i}}{x_i!}\right) \frac{(\sum_{i=17}^{\infty} q_i)^{x_{16}}}{x_{16}!} \\ = \quad & 1192! \frac{(e^{-\lambda}(1 + \lambda + \frac{\lambda^2}{2}))^{x_1}}{x_1!} \left(\prod_{i=2}^{15} \frac{(e^{-\lambda} \frac{\lambda^{i+1}}{(i+1)!})^{x_i}}{x_i!}\right) \frac{(e^{-\lambda} \sum_{i=17}^{\infty} \frac{\lambda^i}{i!})^{x_{16}}}{x_{16}!} \end{aligned} \tag{68}$$

Note that

$$e^{-\lambda} \sum_{i=17}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} \left(\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} - \sum_{i=0}^{16} \frac{\lambda^i}{i!}\right) = e^{-\lambda} \left(e^{\lambda} - \sum_{i=1}^{16} \frac{\lambda^i}{i!}\right) = 1 - e^{-\lambda} \sum_{i=1}^{16} \frac{\lambda^i}{i!} \tag{69}$$

Plugging the result from (69) into (68), we find that, if H_0 is true,

$$f_\theta(x) = 1192! \frac{(e^{-\lambda}(1 + \lambda + \frac{\lambda^2}{2}))^{x_1}}{x_1!} \left(\prod_{i=2}^{15} \frac{(e^{-\lambda} \frac{\lambda^{i+1}}{(i+1)!})^{x_i}}{x_i!}\right) \frac{(1 - e^{-\lambda} \sum_{i=1}^{16} \frac{\lambda^i}{i!})^{x_{16}}}{x_{16}!} \tag{70}$$

for all $x = (x_1, ..., x_{16})$ such that $x_1, ..., x_{16} \geq 0, \in \mathbb{Z}, \sum_{i=1}^{16} x_i = 1192$, and $f_\theta(x) = 0$ otherwise. Now that we have explicitly described $f_\theta(x)$ for all $\theta \in \Theta_0$ (in (70)) and all $\theta \in \Theta$ (in (67)), we can simplify the generalized likelihood ratio from (57). We will start with the numerator. To compute

$$\sup_{\theta \in \Theta} f_\theta(x) = \sup_{(p_1,...,p_{16}) \in \Theta} f_{(p_1,...,p_{16})}(x_1, ..., x_{16})$$

we need to maximize a function in 16 variables, subject to the constraints $\sum_{i=1}^{16} p_i = 1$ and $p_1, ..., p_{16} \geq 0$. We use Lagrange Multipliers. Define

$$g(p_1, ..., p_{16}) = \sum_{i=1}^{16} p_i - 1 = 0 \tag{71}$$

Then, if $\exists p_1, ..., p_{16} \in \Theta$ such that $p_1, ..., p_{16}$ maximizes f over all $\theta \in \Theta$, this $p_1, ..., p_{16}$ must satisfy

$$\nabla_{p_1,...,p_{16}} f = \delta \nabla_{p_1,...,p_{16}} g \tag{72}$$

That is, at any critical point for f, we must have

$$\frac{\partial f_\theta(x)}{\partial p_i} = \delta \frac{\partial g}{\partial p_i} \tag{73}$$

for all $i\{1, ..., 16\}$ and for some $\delta \neq 0$. We can easily compute that

$$\frac{\partial g}{\partial p_i} = \frac{\partial g}{\partial p_i}(p_i) + \frac{\partial g}{\partial p_i}(-1 + \sum_{j \neq i, j \in \{1,...,16\}} p_i) = 1 + 0 = 1 \qquad (74)$$

and

$$\frac{\partial f_\theta(x)}{\partial p_i} = \frac{\partial}{\partial p_i} 1192! \prod_{j=1}^{16} \frac{p_j^{x_j}}{x_j!} = 1192! x_i \frac{p_i^{x_i - 1}}{x_i!} \prod_{j \neq i, j \in \{1,...,16\}} \frac{p_j^{x_j}}{x_j!} = \frac{x_i}{p_i} 1192! \prod_{j=1}^{16} \frac{p_j^{x_j}}{x_j!} = \frac{x_i}{p_i} f_\theta(x) \qquad (75)$$

Plugging the results from (74) and (75) into (73), we find

$$\frac{x_i}{p_i} f_\theta(x) = \delta \implies p_i = \frac{x_i}{\delta} f_\theta(x) \qquad (76)$$

Plugging the result from (76) into (71), we find

$$1 = \sum_{i=1}^{16} p_i = \sum_{i=1}^{16} \frac{x_i}{\delta} f_\theta(x) \implies \frac{\delta}{f_\theta(x)} = \sum_{i=1}^{16} x_i = 1192 \implies \frac{f_\theta(x)}{\delta} = \frac{1}{1192} \qquad (77)$$

Plugging the result from (77) into (76) yields

$$p_i = \frac{x_i}{1192} \qquad (78)$$

This holds for all $i \in \{1, ..., 16\}$, so we know that the only critical point for f on the interior of Θ is

$$(p_1, ..., p_{16}) = (\frac{x_1}{1192}, ..., \frac{x_{16}}{1192}) \qquad (79)$$

Note: Since

$$f_\theta(x) = f_{p_1,...,p_{16}}(x_1, ..., x_n) = 1192! \prod_{i=1}^{16} \frac{p_i^{x_i}}{x_i!} = 0 \qquad (80)$$

for all $p_1, ..., p_{16}$ where $\exists i \in \{1, ..., 16\}$ such that $p_i = 0$. That is, $f_\theta(x) = 0$ for all points on the boundary of Θ. Since $f_\theta(x) \geq 0$ by the definition of a joint probability mass function, this implies the critical point from (79) is indeed the $\theta \in \Theta$ at which $f_\theta(x)$ is *uniquely maximized*. That is

$$\sup_{\theta \in \Theta} f_\theta(x) = \sup_{p_1,...,p_{16} \in \Theta} f_\theta(x) = 1192! \prod_{i=1}^{16} \frac{(\frac{x_i}{1192})^{x_i}}{x_i!} \qquad (81)$$

Now, we can deal with the denominator from (57). Plugging our definition of $f_\theta(x)$, given that H_0 is true, into the denominator from (57) yields

$$\sup_{\theta \in \Theta_0} f_\theta(x) = \sup_{\theta \in \Theta_0} 1192! \frac{(e^{-\lambda}(1 + \lambda + \frac{\lambda^2}{2}))^{x_1}}{x_1!} (\prod_{i=2}^{15} \frac{(e^{-\lambda} \frac{\lambda^{i+1}}{(i+1)!})^{x_i}}{x_i!}) \frac{(1 - e^{-\lambda} \sum_{i=1}^{16} \frac{\lambda^i}{i!})^{x_{16}}}{x_{16}!} \qquad (82)$$

We use a computer and the data from the table to estimate the $\lambda > 0$ which maximizes (82). We find

$$\lambda \approx 8.351 \qquad (83)$$

Plugging the approximate result for λ from (83) into (82), we find

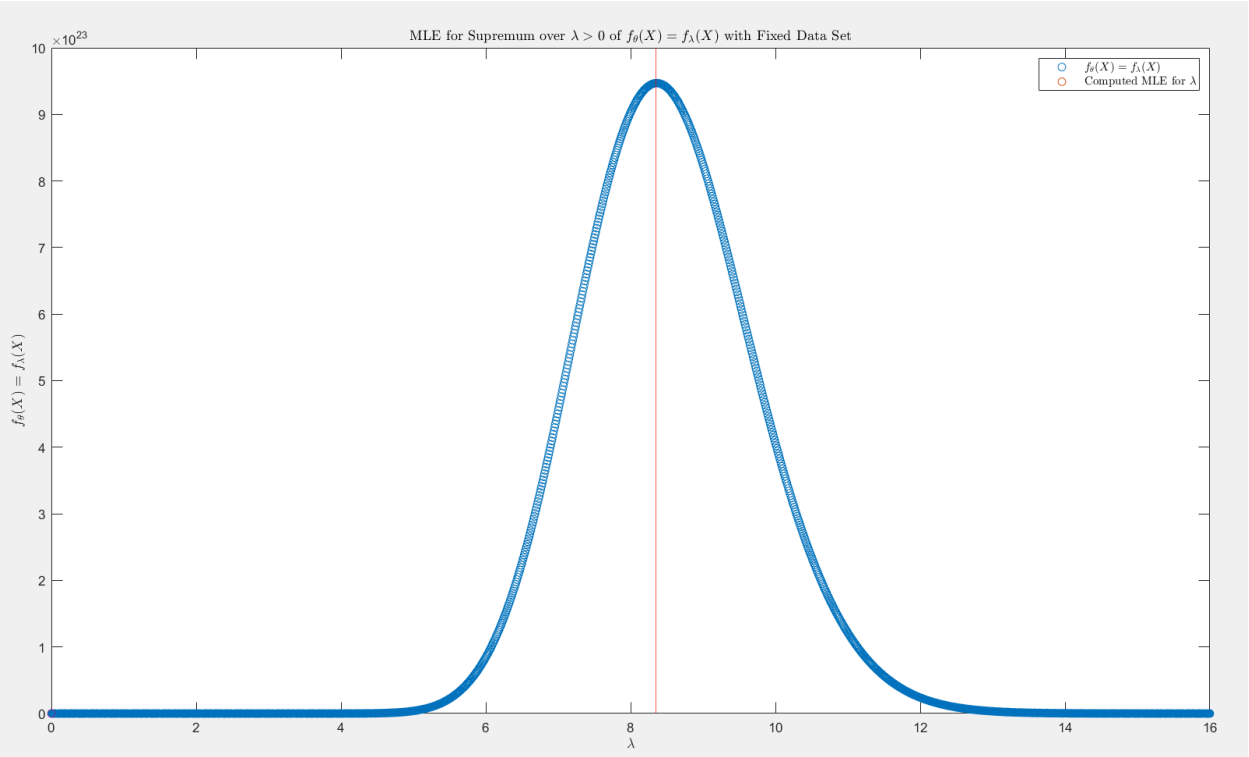
$$\sup_{\theta \in \Theta_0} f_\theta(x) \approx 1192! \frac{(e^{-8.351}(1 + 8.351 + \frac{8.351^2}{2}))^{x_1}}{x_1!} (\prod_{i=2}^{15} \frac{(e^{-8.351} \frac{8.351^{i+1}}{(i+1)!})^{x_i}}{x_i!}) \frac{(1 - e^{-8.351} \sum_{i=1}^{16} \frac{8.351^i}{i!})^{x_{16}}}{x_{16}!} \qquad (84)$$

Combining (84) with (81), we arrive at a final approximation for the generalized likelihood ratio statistic initially presented in (57):

$$\frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)} \approx \frac{1192!\prod_{i=1}^{16}\frac{(\frac{x_i}{1192})^{x_i}}{x_i!}}{1192!\frac{(e^{-8.351}(1+8.351+\frac{8.351^2}{2}))^{x_1}}{x_1!}(\prod_{i=2}^{15}\frac{(e^{-8.351}\frac{8.351^{i+1}}{(i+1)!})^{x_i}}{x_i!})\frac{(1-e^{-8.351}\sum_{i=1}^{16}\frac{8.351^i}{i!})^{x_{16}}}{x_{16}!}}$$

$$= \left(\frac{\frac{x_1}{1192}}{e^{-8.351}(1+8.351+\frac{8.351^2}{2})}\right)^{x_1}\left(\prod_{i=2}^{15}\left(\frac{\frac{x_i}{1192}}{e^{-8.351}\frac{8.351^{i+1}}{(i+1)!}}\right)^{x_i}\right)\left(\frac{\frac{x_{16}}{1192}}{(1-e^{-8.351}\sum_{i=1}^{16}\frac{8.351^i}{i!})}\right)^{x_{16}} \quad (85)$$

We also plot the denominator of the generalized likelihood ratio test statistic as a function of λ to demonstrate how the computed estimate of $\lambda \approx 8.351$ aligns with the distribution of $f_\theta(X) = f_\lambda(X)$:



Note that the $y-axis$ of this plot (i.e. the values of $f_\theta(X)$) are scaled upwards significantly to avoid computational errors arising from large factorials (which evaluate to ∞, or 0 if in the denominator). Both to compute the maximum likelihood estimate of $\lambda \approx 8.351$ and to plot the above figure, we use the following MATLAB code:

```
cols = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16];
freqs = [16, 26, 58, 102, 125, 146, 163, 164, 120, 100, 72, 54, 20, 12, 10, 4];
count = sum(freqs);
x = zeros(16000);
y = zeros(16000);
lambda = 0.0;
blambda = lambda;
max = -1;
for i = 1: 16000
    temp = (((exp(-lambda)*(1+lambda + (lambda^2)/2))^(freqs(1)/200)));
    for j = 2: 15
        temp = temp * (exp(-lambda)*lambda^(j+1))^(freqs(j)/200);
    end
```

```
    s = 0;
    for k = 1: 16
        s = s + (lambda^k)/factorial(k);
    end
    temp  = temp * (1- (exp(-lambda)*s))^(freqs(16)/200);

    if temp > max
        max = temp;
        blambda = lambda;
    end
    lambda = lambda + 0.001;
    y(i) = temp;
    x(i) = lambda;
end
plot(x,y, 'o');
hold on;
xline(blambda, 'Color', 'red');
title({'MLE for Supremum over $\lambda > 0$ of $f_{\theta}(X) = f_{\lambda}(X)$
with Fixed Data Set'}, 'Interpreter', 'latex');
xlabel({'$\lambda$'}, 'Interpreter', 'latex');
ylabel({'$f_{\theta}(X) = f_{\lambda}(X)$'}, 'Interpreter', 'latex');
legend({  '$f_{\theta}(X) = f_{\lambda}(X)$','Computed MLE for $\lambda$'},
'Interpreter', 'latex');
hold off;
```

Now, we will compute the value s of the Pearson's chi-squared statistic S to determine whether we have confidence that the H_0 is true. From lecture, the Pearson's chi-squared statistic is defined to be

$$S := \sum_{i=1}^{16} \frac{\big(X_j - \mathbb{E}_\lambda[X_j]\big)^2}{\mathbb{E}_\lambda[X_j]} \tag{86}$$

and we know that S has a chi-squared distribution with $16 - 1 - 1 = 14$ degrees of freedom. That is, $S \sim \chi^2(14)$, so S has PDF

$$f_S(x) = \frac{1}{2^7\Gamma(7)} x^6 e^{-\frac{x}{2}} \tag{87}$$

Assuming the data set yields a value s of the chi-squared statistic equal to

$$s = \sum_{i=1}^{16} \frac{\big(x_j - \mathbb{E}_\lambda[X_j]\big)^2}{\mathbb{E}_\lambda[X_j]} = \sum_{i=1}^{16} \frac{\big(x_j - 1192p_i\big)^2}{1192p_i} \tag{88}$$

where $p_1, ..., p_{16}$ are defined as in (61), (62), (63) and (64). The last equality follows since $X_i \sim Binomial(1192, p_i)$, so $\mathbb{E}_\lambda[X_i] = 1192p_i$ for all $i \in \{1, ..., 16\}$. We use a computer to approximate the value of (88) with $\lambda = 8.351$, and we find:

$$s \approx 10.9932 \tag{89}$$

Using the PDF from (87), we can directly compute that

$$p = \mathbb{P}_{H_0}(S \geq 10.9932) = \frac{1}{2^7\Gamma(7)} \int_{10.9932}^{\infty} x^6 e^{-\frac{x}{2}} dx \approx 0.6866 \tag{90}$$

The following MATLAB code produced both s and p:

```
s = (freqs(1) - 1192 * exp(-8.351)*(1+8.351+8.351^2/2))^2
/(1192 * exp(-8.351)*(1+8.351+8.351^2/2));
```

```
for i = 2:15
    s= s+ (freqs(i) - 1192*exp(-8.351)*(8.351^(i+1))/(factorial(i+1)))^2
    /(1192*exp(-8.351)*(8.351^(i+1))/factorial(i+1));
end
t = 0;
for i = 1: 16
    t = t + (8.351^i)/factorial(i);
end
t = t * exp(-8.351);
s = s+(freqs(16) - 1192*(1-t))^2 / (1192 *(1-t));
p = 1- chi2cdf(10.9932, 14);
```

Note that p is a p-value corresponding to tests of H_0 of the form

$$C := \{x \in R^n : S \geq c\} \qquad (91)$$

Essentially $p \approx 0.6866$ indicates that, if H_0 is true, there is approximately a 68.66% chance to observe a value s of Pearson's chi-squared statistic S as extreme (large) as $s \approx 10.9932$. Since this probability is so high ($p >> 0.05, p > 0.5$), $p \approx 0.6866$ *does* provide confidence that H_0 is true. That is, the p-value associated with the test described in (91) *does* indicate that the individual counts of alpha particle emissions can be modelled accurately with $i.i.d$ Poisson random variables.

MATH 447: Mathematics of Machine Learning

All assignments in this section were written by Stanislav Minsker, Associate Professor of Mathematics, USC. Solutions to assignments 1 through 5 are provided.

Assignment 1

Read chapter 2, chapter 3 section 1, chapter 4 section 2, chapter 5 section 1 of the textbook "Understanding Machine Learning." Then solve the following problems:

1.

(a) Prove the additive law of probability: for any events A and B,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

(b) Prove the "union bound" by induction: for any $k \geq 2$ and any events $A_1, ..., A_k$,

$$\mathbb{P}(A_1 \cup A_2 \cup ... \cup A_k) \leq P(A_1) + P(A_2) + ... + P(A_k)$$

Solution.
First, we will recall the Axioms of Probability, as they will be essential for both proofs. Define Ω to be the sample space of all possible events:

1. For any event $A \subseteq \Omega$, the probability of A is non-negative:
$$\mathbb{P}(A) \geq 0$$
2. The probability of the entire sample space is 1.
$$\mathbb{P}(\Omega) = 1$$
3. For any events $A_1, A_2, ... \subseteq \Omega$ s.t. $A_i \cap A_j = \emptyset$ for all $1 \leq i, j \in \mathbb{Z}$ where $i \neq j$, the probability of the union of $A_1, A_2, ...$ equals the sum of the probabilities of $A_1, A_2, ...$ That is, for mutually disjoint events $A_1, A_2, ...$, we have
$$\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

(a) To prove that
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$
we want to apply the third axiom of probability. In order to do so, we need to split $(A \cup B)$ into disjoint events. Note that $(A \cup B)$ includes exactly those events in A but not B, in B but not A, and in both A and B. Thus,
$$(A \cup B) = ((A \setminus B) \cup (B \setminus A) \cup (A \cap B))$$
so
$$\mathbb{P}(A \cup B) = \mathbb{P}((A \setminus B) \cup (B \setminus A) \cup (A \cap B)) \qquad (1)$$
Any event in A but not B cannot possibly be in B but not A, nor both A and B. Similarly, any event in B but not A cannot possibly be in A but not B, nor both A and B. Similarly, any event in both A and B cannot possibly be in A but not B, nor B but not A. Thus, we can conclude
$$(A \setminus B) \cap (B \setminus A) = (A \setminus B) \cap (A \cap B) = (B \setminus A) \cap (A \cap B) = \emptyset$$

Since $(A \setminus B)$, $(B \setminus A)$, and $(A \cap B)$ are all mutually disjoint, we can apply the third axiom of probability to (1) to find

$$\mathbb{P}(A \cup B) = \mathbb{P}(A \setminus B) + \mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B) \qquad (2)$$

We can now compute $\mathbb{P}(A \setminus B)$ and $\mathbb{P}(B \setminus A)$ to complete the proof.
Note that, for all $a \in A$ s.t. $a \notin (A \setminus B)$, $a \in B$, so

$$(A \setminus B) \cup (A \cap B) = A \qquad (3)$$

Since $(A \setminus B) \cap (A \cap B) = \emptyset$, we can apply the third axiom of probability to (3) to find

$$\mathbb{P}(A) = \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B) \implies \mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) \qquad (4)$$

Similarly, for all $b \in B$ s.t. $b \notin (B \setminus A)$, $b \in A$, we know

$$(A \cap B) \cup (B \setminus A) = B \qquad (5)$$

Since $(B \setminus A) \cup (A \cap B) = \emptyset$, we can again apply the third axiom of probability to (5) to find

$$\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \setminus A) \implies \mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A \cap B) \qquad (6)$$

Plugging our results from (4) and (6) into (2) yields

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B) \qquad (7)$$

Simplifying (7) yields

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - 2\mathbb{P}(A \cap B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

which completes the proof that

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

(b) We will prove that for any $k \geq 2$ and any events $A_1, ..., A_k$,

$$\mathbb{P}(A_1 \cup A_2 \cup ... \cup A_k) \leq P(A_1) + P(A_2) + ... + P(A_k)$$

by mathematical induction on k.
Base Case: $k = 2$. We want to show $\mathbb{P}(A_1 \cup A_2) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2)$. We could use the additive law of probability in combination with the first axiom of probability. Instead, note that $A_1 \cup A_2$ consists of all events in A_1 and all events in $(A_2 \setminus A_1)$. Thus, we can write

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) \cup \mathbb{P}(A_2 \setminus A_1)$$

By definition of $A_2 \setminus A_1$, for all $a \in A_1$, $a \notin A_2$, so we know $A_1 \cap (A_2 \setminus A_1) = \emptyset$. Thus, we can apply the third axiom of probability to find that

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2 \setminus A_1) \qquad (8)$$

For all $a \in (A_2 \setminus A_1)$, we know $a \in A_2$, so $(A_2 \setminus A_1) \subseteq A_2$, so

$$\mathbb{P}(A_2 \setminus A_1) \leq \mathbb{P}(A_2) \qquad (9)$$

Plugging (9) into (8) yields

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2 \setminus A_1) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2)$$

which completes the proof of the base case.
Inductive Hypothesis: Assume that

$$\mathbb{P}(A_1 \cup ... \cup A_k) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_k)$$

for all $2 \leq k \leq n$.
Inductive Step: Consider $k = n + 1$. We want to show that

$$\mathbb{P}(A_1 \cup ... \cup A_{n+1}) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_{n+1}) \tag{10}$$

Similar to the base case, note that $(A_1 \cup ... \cup A_{n+1})$ consists of all events in $(A_1 \cup ... \cup A_n)$ and all events in $(A_{n+1} \setminus (A_1 \cup ... \cup A_n)$, so we can write

$$\mathbb{P}(A_1 \cup ... \cup A_{n+1}) = \mathbb{P}((A_1 \cup ... \cup A_n) \cup (A_{n+1} \setminus (A_1 \cup ... \cup A_n))) \tag{11}$$

By definition, for all $a \in (A_1 \cup ... \cup A_n)$, $a \notin (A_{n+1} \setminus (A_1 \cup ... \cup A_n))$, so we know

$$((A_{n+1} \setminus (A_1 \cup ... \cup A_n)) \cap (A_1 \cup ... \cup A_n)) = \emptyset$$

Thus, we can apply the third axiom of probability to (11) to find

$$\mathbb{P}(A_1 \cup ... \cup A_{n+1}) = \mathbb{P}(A_{n+1} \setminus (A_1 \cup ... \cup A_n)) + \mathbb{P}(A_1 \cup ... \cup A_n) \tag{12}$$

For all $a \in (A_{n+1} \setminus (A_1 \cup ... \cup A_n))$, $a \in A_{n+1}$, so we know $(A_{n+1} \setminus (A_1 \cup ... \cup A_n)) \subseteq A_{n+1}$, so we have

$$\mathbb{P}(A_{n+1} \setminus (A_1 \cup ... \cup A_n)) \leq \mathbb{P}(A_{n+1}) \tag{13}$$

Also, by the Inductive Hypothesis, we have

$$\mathbb{P}(A_1 \cup ... \cup A_n) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) \tag{14}$$

Plugging the inequalities from (13) and (14) into (12) yields

$$\mathbb{P}(A_1 \cup ... \cup A_{n+1}) = \mathbb{P}(A_{n+1} \setminus (A_1 \cup ... \cup A_n)) + \mathbb{P}(A_1 \cup ... \cup A_n) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n) + \mathbb{P}(A_{n+1})$$

The conclusion that

$$\mathbb{P}(A_1 \cup ... \cup A_k) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_k)$$

follows by induction for all $2 \leq k \in \mathbb{Z}$.

2.

Probability review exercise: recall the notions of joint, marginal and conditional probability density functions (pdf). Solve the following problem: let the joint probability density function of (X, Y) be given by

$$f(y_1, y_2) = \begin{cases} 3y_1 \text{ if } 0 \leq y_2 \leq y_1 \leq 1, \\ 0 \text{ otherwise.} \end{cases} .$$

(a) Find the marginal pdf of X.

(b) Find the conditional pdf of Y given that $X = x$.

(c) Find the conditional expectation of Y given that $X = 1$.

(d*) (bonus) Find the pdf of $X + Y$

Solution.
(a) By definition, to find the marginal PDF of X, we integrate the joint density function $f_{X,Y}(y_1, y_2)$ over all possible values of y_2:

$$f_X(y_1) = \int_{-\infty}^{\infty} f_{X,Y}(y_1, y_2) dy_2 = \int_0^{y_1} 3y_1 dy_2 = 3y_1 y_2 |_0^{y_1} = 3y_1^2$$

This is true for all $0 \leq y_2 \leq y_1 \leq 1$, so the marginal PDF of X is

$$f_X(y_1) = \begin{cases} 3y_1^2 \text{ if } 0 \leq y_1 \leq 1 \\ 0 \text{ otherwise.} \end{cases}$$

(b) By definition, the conditional PDF of Y given $X = x$ is $f_{Y|X=x}(y_2|x) = \frac{f_{X,Y}(x,y_2)}{f_X(x)}$. We know

$$f_X(x) = \begin{cases} 3x^2 \text{ if } 0 \leq x \leq 1 \\ 0 \text{ otherwise.} \end{cases}$$

from part **(a)**, and we are given $f_{X,Y}(x, y_2)$, so we can easily compute that

$$f_{Y|X=x}(y_2|x) = \frac{3x}{3x^2} = \frac{1}{x}$$

This is true for all $0 \leq y_2 \leq x \leq 1$, so the conditional PDF of Y given $X = x$ is

$$f_{Y|X=x}(y_2|x) = \begin{cases} \frac{1}{x} \text{ if } 0 \leq y_2 \leq x \leq 1 \\ 0 \text{ otherwise.} \end{cases}$$

(c) By the definition of conditional expectation, the conditional expectation of Y given that $X = 1$ is

$$\mathbb{E}[Y|X = 1] = \int_{-\infty}^{\infty} y \cdot f_{Y|X=1}(y, 1)dy$$

From part **(b)**, we know that the conditional PDF of Y given $X = 1$ is

$$f_{Y|X=1}(y, 1) = \begin{cases} \frac{1}{1} = 1 \text{ if } 0 \leq y \leq 1 \\ 0 \text{ otherwise.} \end{cases}$$

Thus, we can directly compute that

$$\mathbb{E}[Y|X = 1] = \int_0^1 y dy = \frac{y^2}{2}|_0^1 = \frac{1}{2}$$

So the conditional expectation of Y given $X = 1$ is

$$\mathbb{E}[Y|X = 1] = \frac{1}{2}$$

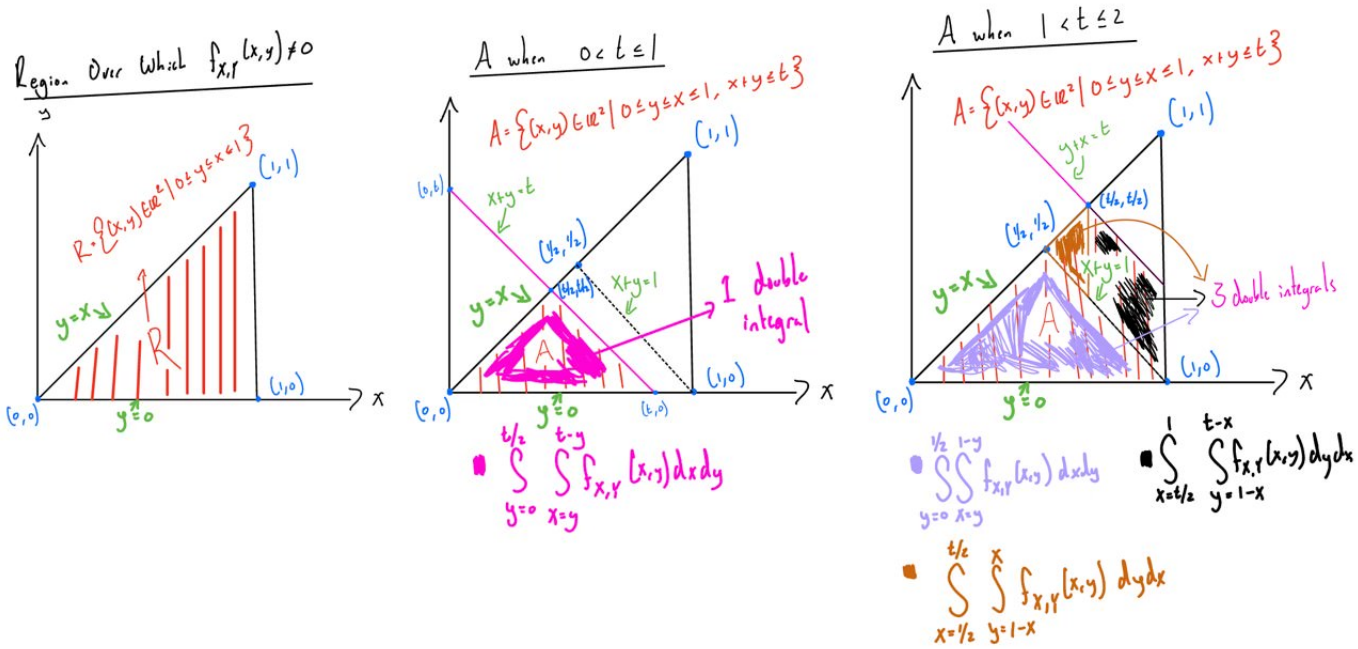
(d) We find the PDF of $X + Y$ by considering the CDF $F_{X+Y}(t)$ and applying the fact that

$$f_{X+Y}(t) = \frac{d}{dt} F_{X+Y}(t)$$

Note that

$$F_{X+Y}(t) = \mathbb{P}(X + Y \leq t) = \int \int_A f_{X,Y}(x, y)dA$$

where $A := \{(x, y) \in \mathbb{R} | 0 \leq y \leq x \leq 1, x + y \leq t\}$. Due to this complicated region of integration A, we need compute $F_{X+Y}(t)$ separately for when $0 < t \leq 1$ and when $1 < t \leq 2$. To help define these cases more clearly, we include a sketch that details how A could look under various t:



This sketch provides the motivation for the following computations. Note that, regardless of t, $A \subseteq R$, where R is the entire region over which $f_{X,Y}(x,y) = 3x$. Therefore, we can plug $f_{X,Y}(x,y) = 3x$ into the integrals established in the sketch to directly compute $F_{X+Y}(t)$ for the two cases of $0 < 1 \leq 1$ and $1 < t \leq 2$. For all $0 < t \leq 1$, we have a singular triangular region over which to integrate. Therefore, we have

$$\begin{aligned} F_{X+Y}(t) = \quad & \mathbb{P}(X+Y \leq t) = \int_{y=0}^{\frac{t}{2}} \int_{x=y}^{t-y} 3x\,dx\,dy = \int_{y=0}^{\frac{t}{2}} \frac{3x^2}{2}\Big|_y^{t-y} dy = \int_{y=0}^{\frac{t}{2}} \frac{3}{2}((t-y)^2 - y^2)dy \\ = \quad & \frac{3}{2}\int_{y=0}^{\frac{t}{2}} t^2 - 2yt\,dy = \frac{3}{2}(t^2 y - y^2 t)\Big|_0^{\frac{t}{2}} = \frac{3}{2}(\frac{t^3}{2} - \frac{t^3}{4}) = \frac{3t^3}{8} \end{aligned} \tag{15}$$

for all $0 < t \leq 1$.
For all $1 < t \leq 2$, we have 2 triangular regions and 1 quadrilateral region over which we must integrate. To compute $F_{X+Y}(t)$ for such t, we integrate over each of these mutually disjoint regions separately and sum the results:

$$F_{X+Y}(t) = \underbrace{\int_{y=0}^{\frac{1}{2}} \int_{x=y}^{1-y} 3x\,dx\,dy}_{1} + \underbrace{\int_{x=\frac{1}{2}}^{\frac{t}{2}} \int_{y=1-x}^{x} 3x\,dy\,dx}_{2} + \underbrace{\int_{x=\frac{t}{2}}^{1} \int_{y=1-x}^{t-x} 3x\,dy\,dx}_{3} \tag{16}$$

Note that integral 1 is just the integral we computed for the $0 < t \leq 1$ case with $t = 1$. Thus, we can plug $t = 1$ into (15) to find

$$\int_{y=0}^{\frac{1}{2}} \int_{x=y}^{1-y} 3x\,dx\,dy = \frac{3(1)^3}{8} = \frac{3}{8} \tag{17}$$

We can directly compute the other two integrals from (16). For integral 2, we have

$$\begin{aligned} \int_{x=\frac{1}{2}}^{\frac{t}{2}} \int_{y=1-x}^{x} 3x\,dy\,dx = \quad & \int_{x=\frac{1}{2}}^{\frac{t}{2}} 3xy\Big|_{1-x}^{x} dx = \int_{x=\frac{1}{2}}^{\frac{t}{2}} (3x^2 - 3x(1-x))dx = \int_{x=\frac{1}{2}}^{\frac{t}{2}} (6x^2 - 3x)dx \\ = \quad & (2x^3 - \frac{3x^2}{2})\Big|_{\frac{1}{2}}^{\frac{t}{2}} = \frac{2t^3}{8} - \frac{3}{8}t^2 - \frac{2}{8} + \frac{3}{8} = \frac{2t^3 - 3t^2 + 1}{8} = \frac{1}{8}(t-1)^2(2t+1) \end{aligned} \tag{18}$$

and for integral 3, we have

$$\begin{aligned}\int_{x=\frac{t}{2}}^{1}\int_{y=1-x}^{t-x} 3xdydx = & \int_{x=\frac{t}{2}}^{1} 3xy|_{1-x}^{t-x}dx = \int_{x=\frac{t}{2}}^{1}(3x(t-x)-3x(1-x))dx = \int_{x=\frac{t}{2}}^{1}(3xt-3x^2-3x+3x^2)dx \\ = & \int_{x=\frac{t}{2}}^{1}(3xt-3x)dx = \int_{x=\frac{t}{2}}^{1}3x(t-1)dx = \frac{3}{2}x^2(t-1)|_{\frac{t}{2}}^{1} = \frac{3}{2}((t-1)-\frac{t^2(t-1)}{4}) \\ = & -\frac{3}{8}(t-1)(t^2-4) \qquad (19)\end{aligned}$$

Plugging the results from (17), (18), and (19) into (16), we find the probability that $X+Y \leq t$, for all $1 < t \leq 2$, is

$$F_{X+Y}(t) = \frac{3}{8} + \frac{1}{8}(t-1)^2(2t+1) - \frac{3}{8}(t-1)(t^2-4)$$

This allows us to fully define the cumulative distribution function of $X+Y$ as

$$F_{X+Y}(t) = \begin{cases} 1 \text{ if } x = t > 2 \\ \frac{3}{8} + \frac{1}{8}(t-1)^2(2t+1) - \frac{3}{8}(t-1)(t^2-4) \text{ if } 1 < t \leq 2 \\ \frac{3t^3}{8} \text{ if } 0 < t \leq 1 \\ 0 \text{ otherwise.} \end{cases}$$

Now, we can apply the fact that

$$f_{X+Y}(t) = \frac{d}{dt}F_{X+Y}(t)$$

to find, for all $t > 2, < 0$

$$f_{X+Y}(t) = \frac{d}{dt}(0) = \frac{d}{dt}(1) = 0$$

for all $0 < t \leq 1$,

$$f_{X+Y}(t) = \frac{d}{dt}\frac{3t^3}{8} = \frac{9t^2}{8}$$

and for all $1 < t \leq 2$,

$$\begin{aligned} f_{X+Y}(t) = & \frac{d}{dt}\big(\frac{3}{8} + \frac{1}{8}(t-1)^2(2t+1) - \frac{3}{8}(t-1)(t^2-4)\big) = \frac{1}{4}(t-1)(2t+1) + \frac{1}{4}(t-1)^2 - \frac{3}{8}(t^2-4) - \frac{3}{4}t(t-1) \\ = & \frac{1}{4}(2t^2-t-1+t^2-2t+1-3t^2+3t) - \frac{3}{8}(t^2-4) = -\frac{3t^2-12}{8}\end{aligned}$$

Thus, we have found that the PDF of $X+Y$ is

$$f_{X+Y}(t) = \begin{cases} -\frac{3t^2-12}{8} \text{ if } 1 < t \leq 2 \\ \frac{9t^2}{8} \text{ if } 0 < t \leq 1 \\ 0 \text{ otherwise.} \end{cases}$$

3.

Suppose that you are tasked with creating an algorithm that, given an image of a handwritten signature of a specific person, decides whether the signature is authentic or whether it was forged by a criminal. Assume that you know in advance that authenticity can be completely determined by the **ratio of signature?s height and width**, measured anywhere from 0.4 to 0.8 with an increment of 0.01 (namely, the ratio for the authentic signatures is either always higher or always lower compared to the forged ones). You have 50 signatures that were randomly taken from various documents that were supposedly signed by this person, and analyzed by an expert who was able to tell originals from the forged ones. Answer the following questions:

(a) What are the instances/observations and the labels in this case? What is the domain set that the instances belong to?
(b) What are the training data in this specific problem, and what is the sample size?
(c) What is the base class that you are going to use, and what is its size/cardinality?
Be as specific as possible; your answer should be consistent with your answer to question (a).
(d) Is this an example of realisable or agnostic learning? Justify your answer.
(e) Assume that you want to find a classifier that makes a mistake in at most 5% of the cases. Estimate the probability that the Empirical Risk Minimization algorithm based on the sample size you determined will produce a classifier of such quality (hint: use the bound we proved in class when showing that finite hypotheses classes are PAC learnable).
(f) What is the size of the training data you would need to construct a classifier that is 95% accurate on the whole population with probability at least 99% (in other words, for 99% of the possible training samples)?

Solution.
(a) The instances/observations in this case are the height:width ratios X of signatures of the person in question. The labels $Y \in \{\pm 1\}$ are the authenticity or forged status of the signatures, such that, for a given instance (height:width ratio) X, it's corresponding label Y satisfies

$$Y = \begin{cases} +1 \text{ if } X \text{ denotes an authentic signature} \\ -1 \text{ otherwise.} \end{cases}$$

Since the height and width ratios are guaranteed to be between 0.4 and 0.8, measured in increments of 0.01, we know that X must belong to the set

$$S := \{0.40, 0.41, 0.42, ..., 0.78, 0.79, 0.80\}$$

That is, for all X, $X \in S$, so S is our domain set. Note that there are exactly 41 numbers between 0.40 and 0.80 (inclusive), so $|S| = 41$.

(b) In this specific problem, the training data are

$$\mathbb{X} = \{(X_1, Y_1), (X_2, Y_2), ..., (X_{50}, Y_{50})\}$$

where X_i denotes the height:weight ratio of the ith randomly selected signature, and

$$Y_i = \begin{cases} +1 \text{ if the } i\text{th signature is authentic} \\ -1 \text{ otherwise.} \end{cases}$$

Since we have 50 signatures, we have 50 pairs $(X_1, Y_1), ..., (X_{50}, Y_{50})$ of height:width ratios and authenticity labels, so our sample size is $n = 50$.

(c) We know that the height:width ratio for authentic signatures is either always higher or always lower than the forged ones. To figure out the direction of this inequality, we can define

$$Y_{max} := Y_i \text{ s.t. } X_i = \max\{X_1, ..., X_{50}\}$$

and define our base class based on the value of Y_{max}. If $Y_{max} = +1$, we know the sample signature with the largest height:width ratio is authentic, so we can assume that the authentic signatures always have larger ratios than the forged ones. If $Y_{max} = -1$, we know the sample signature with the largest height:width ratio is forged, so we can assume that the authentic signatures always have smaller ratios than the forged ones. In either case, we have the base class

$$G := \{g_t : S \to \{\pm 1\} | t \in S\}$$

where

$$g_t(x) := \begin{cases} \begin{cases} +1 \text{ if } x \geq t \\ -1 \text{ otherwise.} \end{cases} & \text{if } Y_{max} = +1 \\ \begin{cases} +1 \text{ if } x \leq t \\ -1 \text{ otherwise.} \end{cases} & \text{otherwise.} \end{cases}$$

Regardless of the value of Y_{max} the cardinality of the base class is

$$|G| = |S| = 41$$

as, for each $t \in S$, $\exists$ exactly 1 $g_t \in G$ which maps instances to labels depending on that t. Thus, there is a 1-to-1 ratio between the possible height:width ratios in S and the possible classifier functions in G. This ensures the consistency of the size of the base class with the size of the domain of the instances.
Note: If we eliminate the Y_{max} variable to create a base class G that doesn't depend on the training data, we could define

$$G := \{g_t', g_t" : S \to \{\pm 1\} | t \in S\}$$

where

$$g_t'(x) := \begin{cases} +1 \text{ if } x \geq t \\ -1 \text{ otherwise.} \end{cases} \qquad g_t"(x) := \begin{cases} +1 \text{ if } x \leq t \\ -1 \text{ otherwise.} \end{cases}$$

In this case, regardless of the training data, we would have a base class G with cardinality $|G| = 2|S| = 2 * 41 = 82$, as, for all $t \in S$, $\exists g_t', g_t" \in G$. This 2:1 ratio between $|G|$ and $|S|$ demonstrates consistency between the size of the base class and the size of the instance domain.

(d) This is an example of realizable learning. This follows from the assumption that "the ratio for the authentic signatures is either always higher or always lower compared to the forged ones" and the restriction of instance height:width ratios to the 41 increments of 0.01 between 0.4 and 0.8 that make up S. Since the authentic ratios are always higher or lower than the forged ones, we know there exists an ideal threshold t^* such that all ratios $X < t^*$ belong to one label and all ratios $X \geq t^*$ belong to the other label. Moreover, since this t^* must be one of the 41 increments of 0.01 between 0.4 and 0.8 (inclusive), and S consists of all of these increments, we know that $t^* \in S$. Since $\exists g_t \in G$ for all $t \in S$, we know that $\exists g_{t^*} \in G$ s.t.

$$\mathbb{P}(g_{t^*}(X) = Y) = 1$$

where X is any height:width ratio in S and $Y \in \{\pm 1\}$ is its corresponding authenticity label. Since the ideal, perfect classifier g_{t^*} is guaranteed to be a part of our base class G, we know this is an example of realizable learning.
Note: In the case where we discard Y_{max} and have $|G| = 82$, the example is still realizable, as $g_{t^*} \in G$ is still guaranteed by the same line of reasoning.

(e) Let $\hat{g}_{50}$ be a classifier function $g \in G$ satisfying Empirical Risk Minimization. Let $L(\hat{g}_{50}) =$ the probability that $\hat{g}_{50}$ incorrectly labels a randomly selected signature. Then the probability we want to estimate is

$$\mathbb{P}(L(\hat{g}_{50}) \leq 0.05) = 1 - \mathbb{P}(L(\hat{g}_{50}) \geq 0.05) \qquad (20)$$

From lecture, since $|G| = 41 < \infty$, we know

$$\mathbb{P}(L(\hat{g}_{50}) \geq \varepsilon) \leq |G| e^{-\varepsilon n}$$

Since $n = 50$ for our training data and we want a mistake-rate of no more than $\varepsilon = 0.05$, we can estimate that

$$\mathbb{P}(L(\hat{g}_{50}) \geq 0.05) \leq 41 \cdot e^{-0.05 \cdot 50} = 41 \cdot e^{-2.5} \approx 3.365$$

Since this upper bound for $\mathbb{P}(L(\hat{g}_{50}) \geq 0.05)$ is significantly greater than 1, and we know no event has probability > 1, we can estimate that

$$\mathbb{P}(L(\hat{g}_{50}) \geq 0.05) \approx 1 \qquad (21)$$

Plugging (21) into (20) yields

$$\mathbb{P}(L(\hat{g}_{50}) \leq 0.05) \approx 1 - 1 = 0$$

Thus, with training data of size $n = 50$, the probability that the Empirical Risk Minimization algorithm will produce a classifier that makes a mistake in at most 5% of cases is $\approx 0\%$. This suggests we need a significantly larger sample size in order to use Empirical Risk Minimization to consistently produce classifiers that rarely make mistakes.

Note: In the case where we discard Y_{max} for a base class G of size $|G| = 82$, we can still apply the same upper bound estimate on $\mathbb{P}(L(\hat{g}_{50}) \geq 0.05)$ since $|G| = 82 < \infty$. Doing so yields

$$\mathbb{P}(L(\hat{g}_{50}) \leq 0.05) = 81e^{-2.5} \approx 6.731$$

Once again, since this probability is significantly larger than 1, we can estimate that

$$\mathbb{P}(L(\hat{g}_{50}) \leq 0.05) \approx 1$$

which implies that

$$\mathbb{P}(L(\hat{g}_{50}) \leq 0.05) \approx 1 - 1 = 0$$

Once again, with a sample size of only $n = 50$, it is very unlikely that the Empirical Risk Minimization algorithm will produce a classifier that makes mistakes at most 5% of the time.

(f) We want to find an n such that

$$\mathbb{P}(L(\hat{g}_{50}) \geq 0.05) \leq 0.01$$

Since $|G| = 41 < \infty$, we know from lecture that

$$\mathbb{P}(L(\hat{g}_{50}) \geq \varepsilon) \leq |G|e^{-\varepsilon n} \leq \delta \iff n \geq \frac{1}{\varepsilon} ln(\frac{|G|}{\delta})$$

Thus, setting $\delta = 1 - 0.99 = 0.01$ and $\varepsilon = 0.05$, we find

$$n \geq \frac{1}{0.05} ln(\frac{41}{0.01}) = 20ln(4100) \approx 166.375$$

This means, for our base class G of size $|G| = 41$, we must have training data of sample size $n \geq 167$ to guarantee that the Empirical Risk Minimization algorithm will produce a classifier that is at least 95% accurate on at least 99% of possible training samples.

Note: If we discard Y_{max} and consider the base class G of size $|G| = 82$, we can still apply the same lower bound on n with the same epsilon and δ since $|G| = 82 < \infty$. In this case, we can guarantee

$$\mathbb{P}(L(\hat{g}_{50}) \geq 0.05) \leq 0.01$$

for

$$n \geq 20ln(8200) \approx 180.238$$

Thus, for a base class of size $|G| = 82$, we would need to have a sample of size at least $n \geq 181$ to guarantee that the Empirical Risk Minimization algorithm will produce a classifier that is 95% accurate on the population with probability at least 99%. Note that, since there are more possible classifiers for our algorithm to choose from with a bigger G, and still only one ideal classifier $g_{t^*} \in G$, we require a larger sample to guarantee the algorithm finds an equally accurate classifier on the same percentage of possible training samples.

Assignment 2

Read chapters 3, 4 and 5 of the textbook "Understanding Machine Learning." Then do the following problems:

1.

Let (X, Y) be a pair of random variables and $\eta(x) = \mathbb{E}[Y|X = x]$ is the conditional expectation. Prove that for any function $g(X)$,

$$\mathbb{E}(Y - \eta(X))g(X) = 0$$

You may use a hint given in the class notes.

Solution.
We substitute X for W and Y for Z, then apply the hint from the class notes that the minimum of the function

$$F(t) = \mathbb{E}[(Y - (\eta(X) + t \cdot g(X)))^2] \qquad (1)$$

is attained at $t = 0$ for any function g. Since $F(t)$ is a concave-up quadratic in t, we know

$$f'(t) = 0 \iff F(t) \text{ is minimized.}$$

Applying the hint that $F(t)$ is minimized at $t = 0$, we find

$$f'(0) = 0 \qquad (2)$$

Using the chain rule, we can directly compute that

$$\begin{aligned} f'(t) = & \quad \frac{d}{dt}F(t) = \frac{d}{dt}\mathbb{E}[Y - (\eta(X) + t \cdot g(X))]^2 \\ = & \quad 2 \cdot \mathbb{E}[\big(Y - (\eta(X) + t \cdot g(X))\big) \cdot \frac{d}{dt}(t \cdot g(X))] \\ = & \quad 2\mathbb{E}[\big(Y - (\eta(X) + t \cdot g(X))\big)g(X)] \qquad (3) \end{aligned}$$

Plugging $t = 0$ into (3) yields

$$f'(0) = 2\mathbb{E}[\big(Y - (\eta(X) + 0 \cdot g(X))\big)g(X)] = 2\mathbb{E}[\big(Y - \eta(X)\big)g(X)] \qquad (4)$$

Comparing (4) with (2), we find

$$0 = f'(0) = 2\mathbb{E}[\big(Y - \eta(X)\big)g(X)] \qquad (5)$$

Since $2 \neq 0$, (5) directly implies that

$$\mathbb{E}[\big(Y - \eta(X)\big)g(X)] = 0$$

Since the hint we used to derive this conclusion holds for all functions g, this completes the proof that

$$\mathbb{E}[(Y - \eta(X))g(X)] = 0$$

for any function $g(X)$.

2.

Assume that you draw 2 cards at random from a deck of 36 cards. Let X take values 1,2,3 if a pair contains no numbered cards (meaning 6,7,8,9,10), 1 numbered card, or 2 numbered cards respectively. Assume that $Y \in \{+1, -1\}$ is such that

$$\mathbb{P}(Y = 1|X = x) = \begin{cases} \frac{1}{3} \text{ if } x = 1 \\ \frac{2}{5} \text{ if } x = 2 \\ \frac{3}{4} \text{ if } x = 3 \end{cases}$$

(a) Find the generalization error of a classifier $h(x) = \begin{cases} -1 \text{ if } x = 2 \\ 1 \text{ otherwise} \end{cases}$

(b) Find the Bayes classifier.

Solution.

(a) By definition, the generalization error of a classifier $h(x)$ is

$$L(h) = \mathbb{P}(Y \neq h(X)) \tag{6}$$

Note that $(X = 1)$, $(X = 2)$, and $(X = 3)$ are three mutually disjoint events whose union is

$$(X = 1) \cup (X = 2) \cup (X = 3) = \Omega$$

where Ω is the sample space for X. Thus, we can apply the Law of Total Probability to (6) to find

$$L(h) = \mathbb{P}(Y \neq h(X), X = 1) + \mathbb{P}(Y \neq h(X), X = 2) + \mathbb{P}(Y \neq h(X), X = 3) \tag{7}$$

Applying the definition of conditional probability to (7) yields

$$L(h) = \mathbb{P}(Y \neq h(X)|X = 1)\mathbb{P}(X = 1)+\mathbb{P}(Y \neq h(X)|X = 2)\mathbb{P}(X = 2)+\mathbb{P}(Y \neq h(X)|X = 3)\mathbb{P}(X = 3) \tag{8}$$

Now, we just have to individually calculate the probabilities in (8) for the given $h(x)$.
Note that

$$\mathbb{P}(Y \neq h(X)|X = 1) = \mathbb{P}(Y = -1, h(X) = 1|X = 1) + \mathbb{P}(Y = 1, h(X) = -1|X = 1) \tag{9}$$

We know that $X = 1 \implies h(X) = 1$, so $\mathbb{P}(Y = 1, h(X) = -1|X = 1) = 0$ and $\mathbb{P}(Y = -1, h(X) = -1|X = 1) = \mathbb{P}(Y = -1|X = 1)$. The latter equality follows from the fact that $h(X) = 1$ is always true when $X = 1$, so $Y \neq h(X)|X = 1$ can only happen when $Y = -1$. This knowledge allows us to rewrite (9) as

$$\mathbb{P}(Y \neq h(X)|X = 1) = \mathbb{P}(Y = -1|X = 1) + 0 = \mathbb{P}(Y = -1|X = 1) \tag{10}$$

We are given $\mathbb{P}(Y = 1|X = 1)$, so we can easily compute that

$$\mathbb{P}(Y = -1|X = 1) = 1 - \mathbb{P}(Y = 1|X = 1) = 1 - \frac{1}{3} = \frac{2}{3} \tag{11}$$

Plugging (11) into (10) yields

$$\mathbb{P}(Y \neq h(X)|X = 1) = \frac{2}{3} \tag{12}$$

Similarly,

$$\mathbb{P}(Y \neq h(X)|X = 2) = \mathbb{P}(Y = 1, h(X) = -1|X = 2) + \mathbb{P}(Y = -1, h(X) = 1|X = 2) \tag{13}$$

We know that $X = 2 \implies h(X) = -1$, so $\mathbb{P}(Y = -1, h(X) = 1|X = 2) = 0$ and $\mathbb{P}(Y = 1, h(X) = -1|X = 2) = \mathbb{P}(Y = 1|X = 2)$ with the latter equality again following from the fact that $h(X) = -1$ is always true for $X = 2$, so $Y \neq h(X)|X = 2$ can only happen when $Y = 1$. This allows us to rewrite (13) as

$$\mathbb{P}(Y \neq h(X)|X = 2) = \mathbb{P}(Y = 1|X = 2) + 0 = \mathbb{P}(Y = 1|X = 2) \tag{14}$$

Plugging the given value for $\mathbb{P}(Y = 1|X = 2)$ into (14) yields

$$\mathbb{P}(Y \neq h(X)|X = 2) = \frac{2}{5} \tag{15}$$

Similarly,

$$\mathbb{P}(Y \neq h(X)|X=3) = \mathbb{P}(Y=-1, h(X)=1|X=3) + \mathbb{P}(Y=1, h(X)=-1|X=3) \tag{16}$$

We know that $X=3 \implies h(X)=1$, so $\mathbb{P}(Y=1, h(X)=-1|X=3)=0$ and $\mathbb{P}(Y=-1, h(X)=1|X=3) = \mathbb{P}(Y=-1|X=3)$, with the latter equality again following from the fact that $h(X)=1$ is always true for $X=3$, so $Y \neq h(X)|X=3$ can only happen when $Y=-1$. We can now rewrite (16) as

$$\mathbb{P}(Y \neq h(X)|X=3) = \mathbb{P}(Y=-1|X=3) + 0 = \mathbb{P}(Y=-1|X=3) \tag{17}$$

Since we are given $\mathbb{P}(Y=1|X=3)$, we can easily compute that

$$\mathbb{P}(Y=-1|X=3) = 1 - \mathbb{P}(Y=1|X=3) = 1 - \frac{3}{4} = \frac{1}{4} \tag{18}$$

Plugging (18) into (17) yields

$$\mathbb{P}(Y \neq h(X)|X=3) = \frac{1}{4} \tag{19}$$

Now, we just need to compute $\mathbb{P}(X=x)$ for $x \in \{1,2,3\}$. Since we assume the pair of cards is drawn at random from a deck of 36 cards, we know all pairs of cards are equally likely. For any sample space Ω in which $\mathbb{P}(a) = \mathbb{P}(b)$ for all $a, b \in \Omega$, we know that for any $X \subseteq \Omega$, we have

$$\mathbb{P}(X) = \frac{|X|}{|\Omega|} \tag{20}$$

Our sample space Ω consists of all 2-card pairs from a deck of 36 cards. There are exactly $\binom{36}{2}$ ways to choose such a pair. Thus, we have

$$|\Omega| = \binom{36}{2} \tag{21}$$

Note that, in a deck of 36 cards containing all 4 suits of cards (6,7,8,9,10,J,Q,K,A), exactly $5 \cdot 4 = 20$ of the cards are numbered cards while the remaining $4 \cdot 4 = 16$ are not numbered cards. For $X=1$, we need to select 0 numbered cards. Thus, we must select 2 cards from the 16 non-numbered cards in the deck. There are exactly $\binom{16}{2}$ ways to do this, so

$$|X=1| = \binom{16}{2} \tag{22}$$

Plugging (22) and (21) into (20) yields

$$\mathbb{P}(X=1) = \frac{\binom{16}{2}}{\binom{36}{2}} \approx 0.1905 \tag{23}$$

Similarly, for $X=2$, we need to select 1 of the 20 numbered cards and one of the 16 non-numbered cards. There are exactly $\binom{20}{1}\binom{16}{1} = 20 \cdot 16 = 320$ ways to do this, so

$$|X=2| = 320 \tag{24}$$

Plugging (24) and (21) into (20) yields

$$\mathbb{P}(X=2) = \frac{320}{\binom{36}{2}} \approx 0.5079 \tag{25}$$

Similarly, for $X=3$, we need to select 2 numbered cards from the 20 numbered cards in the deck. There are exactly $\binom{20}{2}$ ways to do this, so

$$|X=3| = \binom{20}{2} \tag{26}$$

Plugging (26) and (21) into (20) yields

$$\mathbb{P}(X=3)=\frac{\binom{20}{2}}{\binom{36}{2}}\approx 0.3016 \qquad (27)$$

Now, we simply plug (27), (25), (23), (19), (15), and (12) into (8) to find

$$L(h)=\frac{2}{3}\frac{\binom{16}{2}}{\binom{36}{2}}+\frac{2}{5}\frac{320}{\binom{36}{2}}+\frac{1}{4}\frac{\binom{20}{2}}{\binom{36}{2}}\approx 0.4056$$

Thus, the generalization error of $h(x)$ is

$$L(h)=\mathbb{P}(Y\neq h(x))\approx 0.4056$$

so there is approximately a 40.56% chance that h(x) will misclassify a given instance X.

(b) Now, we will find the Bayes classifier. Define $\eta(x)\coloneqq\mathbb{E}[Y|X=x]$. By definition, the Bayes classifier is

$$g_*(x)=\begin{cases}+1 \text{ if } \eta(x)\geq 0\\ -1 \text{ if } \eta(x)<0\end{cases} \qquad (28)$$

Applying the definition of conditional expectation, we find

$$\eta(x)=\mathbb{E}[Y|X=x]=\sum_{y\in\{-1,+1\}} y\mathbb{P}(Y=y|X=x)=-\mathbb{P}(Y=-1|X=x)+\mathbb{P}(Y=1|X=x) \qquad (29)$$

Plugging $x=1$ into (29) yields

$$\eta(1)=-(1-\mathbb{P}(Y=1|X=1))+\mathbb{P}(Y=1|X=1)=-(1-\frac{1}{3})+\frac{1}{3}=\frac{-2}{3}+\frac{1}{3}=-\frac{1}{3} \qquad (30)$$

Since $\eta(1)=-\frac{1}{3}<0$, we know from (28) that

$$g_*(1)=-1 \qquad (31)$$

Plugging $x=2$ into (29) yields

$$\eta(2)=-(1-\mathbb{P}(Y=1|X=2))+\mathbb{P}(Y=1|X=2)=-(1-\frac{2}{5})+\frac{2}{5}=-\frac{3}{5}+\frac{2}{5}=-\frac{1}{5} \qquad (32)$$

Since $\eta(2)=-\frac{1}{5}<0$, we know from (28) that

$$g_*(2)=-1 \qquad (33)$$

Plugging $x=3$ into (29) yields

$$\eta(3)=-(1-\mathbb{P}(Y=1|X=3))+\mathbb{P}(Y=1|X=3)=-(1-\frac{3}{4})+\frac{3}{4}=-\frac{1}{4}+\frac{3}{4}=\frac{1}{2} \qquad (34)$$

Since $\eta(3)=\frac{1}{2}\geq 0$, we know from (28) that

$$g_*(3)=1 \qquad (35)$$

Combining (31), (33), and (35) yields

$$g_*(x)=\begin{cases}+1 \text{ if } x=3\\ -1 \text{ if } x\in\{1,2\}\end{cases} \qquad (36)$$

By definition, (36) is the Bayes classifier for this problem.

3.

Let S be a discrete but possibly infinite subset of $\mathbb{R}$, and consider the infinite family of binary classifiers G that consists of all functions $g_z(x) = \begin{cases} 1 \text{ if } x = z \\ -1 \text{ if } x \neq z \end{cases}$ indexed by $z \in S$, and also includes the classifier $g^-(x)$ that is identically equal to -1. Assume the realizable learning scenario and let $(X_1, Y_1), ..., (X_n, Y_n)$ be the training data.

(a) Describe the ERM algorithm adapted to this specific case

(b) Show that the class G is PAC learnable despite being infinite.

Hint for part (2): let $\hat{g}$ be the output of the ERM algorithm. We want to estimate $\mathbb{P}(L(\hat{g}) > \varepsilon)$ for some $\varepsilon > 0$. Consider 3 possibilities: (a) the perfect classifier is g^-; (b) the perfect classifier is $g_z$for some $z \in S$ and the distribution of X is such that $\mathbb{P}(X = z) \leq \varepsilon$; (c) the perfect classifier is g_z for some $z \in S$ and the distribution of X is such that $\mathbb{P}(X = z) > \varepsilon$.

Solution.

(a) By the definition of Empirical Risk Minimization, and since we assume the realizable learning scenario, we know that the ERM algorithm will pick a $\hat{g} \in G$ s.t. $\hat{g}(X_i) = Y_i$ for all $i \in \{1, ..., n\}$.

Note: Since we assumed the realizable learning scenario, we know $\exists g_* \in G$ s.t. $\mathbb{P}(Y = g_*(X)) = 1$ (i.e. the perfect/ideal classifier). If this classifier is some g_{z_i} where $z_i \in S$, then we know $\mathbb{P}(Y = g_{z_j}(z_i)) = 0$ for all $z_j \in S$ s.t. $z_j \neq z_i$ and $\mathbb{P}(Y = g^-(z_i)) = 0$ since we know $g_{z_i}(z_i) = 1$ is the correct classification of z_i while $g_{z_j}(z_i) = g^-(z_i) = -1$ is the incorrect classification of z_i. Similarly, if the ideal classifier $g_* = g^-$, then we know $\mathbb{P}(g_z(z) = Y) = 0$ for all $z \in S$ as $g^-(z) = 0$ is the correct classification of each $z \in S$ while $g_z(z) = 1$ is the incorrect classification. Thus, regardless of the true value of g_*, we know that only 1 perfect classifier exists.

Since we know there exists only 1 perfect classifier g_*, and no classifier is perfect if $Y_i = Y_j = 1$, $X_i \neq X_j$, we know that

$$|\{i \in \{1, ..., n\} | Y_i = 1\}| \in \{0, 1\}$$

If $|\{i \in \{1, ..., n\} | Y_i = 1\}| = 1$, we know there exists a unique $i \in \{1, ..., n\}$ s.t. $g_*(X_i) = 1$. Since $g^-(X_i) = -1$ and $g_z(X_i) = -1$ for all $z \in S$ s.t. $z \neq X_i$, g_{X_i} is the *only* $g \in G$ s.t. $g(X_i) = Y_i$ for all $i \in \{1, ..., n\}$. Thus, if $\exists i \in \{1, ..., n\}$ s.t. $g_*(X_i) = Y_i = 1$, the ERM algorithm will output $\hat{g} = g_{X_i}$, as this is the only classifier consistent with the training data.
Note: This explanation assumes that $X_i \neq X_j$ for all $i \neq j \in \{1, ..., n\}$. However, if we are given duplicate training instances, since we assume all training data is labeled correctly, we know these duplicates will have the same label. If $(X_i, Y_i) = (X_j, Y_j)$ for some $j \neq i$, then (X_j, Y_j) gives us no additional information about the true distribution of X. Thus, we can simply ignore these duplicates to ensure $|\{i \in \{1, ..., n\} | Y_i = 1\}| \in \{0, 1\}$ and apply the same logic. If we have duplicates $(X_i, 1)$ and $(X_j, 1)$, then $g_{X_i} = g_{X_j}$, so the ERM algorithm can still always output $\hat{g} = g_{X_i}$ in this case. In general, if $|\{i \in \{1, ..., n\} | Y_i = 1\}| > 1$, for all $i, j \in \{i \in \{1, ..., n\} | Y_i = 1\}$, $g_{X_i} = g_{X_j}$, so ignoring duplicates does not change the output of the ERM algorithm.

If $|\{i \in \{1, ..., n\} | Y_i = 1\}| = 0$, the only other possible case, then we know $Y_i = -1$ for all $i \in \{1, ..., n\}$. Since $g^-(x) = -1$ for all $x \in \mathbb{R}$, we know $g^-(X_i) = -1 = Y_i$ for all $i \in \{1, ..., n\}$, so g^- is consistent with the training data. Also, by definition, for all g_z s.t. $z \in S$, $z \notin \{X_1, ..., X_n\}$, we know

$g_z(X_i) = -1 = Y_i$ for all $i \in \{1, ..., n\}$, so all such g_z are consistent with the training data. Thus, if $|\{i \in \{1, ..., n\} | Y_i = 1\}| = 0$, we know the ERM algorithm will output some

$$\hat{g} = g \in \{g^-, g_z | z \notin \{X_1, ..., X_n\}\} \subseteq G$$

as $\{g^-, g_z | z \notin \{X_1, ..., X_n\}\}$ consists of all classifiers in G that are consistent with the training data.

Since ignoring duplicate instances doesn't change the output of the ERM algorithm, we can assume $|\{i \in \{1, ..., n\} | Y_i = 1\}| \in \{0, 1\}$, allowing us to describe all possibilities for the output of the ERM algorithm with

$$\hat{g} = \begin{cases} g_{X_i} & \text{if} \quad |\{i \in \{1, ..., n\} | Y_i = 1\}| = 1 \\ g \in \{g^-, g_z | z \notin \{X_1, ..., X_n\}\} & \text{if} \quad |\{i \in \{1, ..., n\} | Y_i = 1\}| = 0 \end{cases}$$

which completes our description of the ERM algorithm adapted to this case.

(b) First, we will slightly modify the ERM algorithm described in part (a) to make it more deterministic. This will drastically simplify the following probability calculations. We can do this since, to prove PAC learnability of a class G, we just need to present *one* algorithm under which G is PAC learnable.
Once again, since ignoring duplicate instances doesn't change the output of the ERM algorithm, we can assume $|\{i \in \{1, ..., n\} | Y_i = 1\}| \in \{0, 1\}$. We only change the case when $|\{i \in \{1, ..., n\} | Y_i = 1\}| = 0$. In such a case, we know that $Y_i = -1$ for all $i \in \{1, .., n\}$, so we know $g^-(X_i) = Y_i$ for all $i \in \{1, ..., n\}$ since $g^-(x) := -1$ for all $x \in \mathbb{R}$. Thus, when $|\{i \in \{1, ..., n\} | Y_i = 1\}| = 0$, we have

$$\mathbb{P}(g^-(X_i) \neq Y_i) = 0 \qquad (37)$$

So g^- has empirical risk of $L_n(g^-) = 0$. Since 0 is the minimum possible empirical risk, we know $L_n(g) \geq L_n(g^-)$ for all $g \in G$. This allows us to *always* choose $\hat{g} = g^-$ when $|\{i \in \{1, ..., n\} | Y_i = 1\}| = 0$ while still minimizing the empirical risk of $\hat{g}$ across all $g \in G$. Thus, we can describe all possible outputs of our modified ERM algorithm with

$$\hat{g}_m := \begin{cases} g_{X_i} \text{ if } |\{i \in \{1, ..., n\} | Y_i = 1\}| = 1 \\ g^- \text{ if } |\{i \in \{1, ..., n\} | Y_i = 1\}| = 0 \end{cases} \qquad (38)$$

where $\hat{g}_m$ is the classifier from G which our modified algorithm outputs. Now, given any training data $(X_1, Y_1), ..., (X_n, Y_n)$, we can determine the modified ERM algorithm's output with certainty.

To prove PAC learnability, we need to

(i) Present an algorithm A with output $\hat{g}$

(ii) Find a sample complexity function $n(\varepsilon, \delta)$ s.t.

$$\mathbb{P}(L(\hat{g}) > \varepsilon) < \delta \qquad \forall (\varepsilon, \delta) \in (0, 1)^2$$

for any sample of size $n \geq n(\varepsilon, \delta)$, $n \in \mathbb{Z}$.

We already completed part (i) by presenting the modified ERM algorithm with output $\hat{g}_m$. Now, we will complete part (ii) by applying the given hint and splitting the scenario into 3 cases. Let $A_1 =$ the event that the perfect classifier is g^-. Let $A_2 =$ the event that the perfect classifier is g_z for some $z \in S$ and the distribution of X is s.t. $\mathbb{P}(X = z) \leq \varepsilon$. Let $A_3 =$ the event that the perfect classifier is g_z for some $z \in S$ and the distribution of X is s.t. $\mathbb{P}(X = z) > \varepsilon$. Then we have

$$\mathbb{P}(L(\hat{g}_m) > \varepsilon) = \mathbb{P}(L(\hat{g}_m) > \varepsilon | A_1)\mathbb{P}(A_1) + \mathbb{P}(L(\hat{g}_m) > \varepsilon | A_2)\mathbb{P}(A_2) + \mathbb{P}(L(\hat{g}_m) > \varepsilon | A_3)\mathbb{P}(A_3) \qquad (39)$$

Case 1: The perfect classifier is g^-. By the definition of a perfect classifier, we have $\mathbb{P}(Y \neq g^-(X) | A_1) = 0$, and we know $\mathbb{P}(Y_i \neq g^-(X_i)) = 1$ for all $Y_i = 1$ since $g^-(x) := -1$ for all $x \in \mathbb{R}$, so a perfect classifier

of g^- implies that $|\{i \in \{1,...,n\}|Y_i = 1\}| = 0$. In this case, the output of our modified ERM algorithm is $\hat{g}_m := g^-$. Thus, if g^- is the perfect classifier, our modified ERM algorithm will always output the perfect classifier. We can easily verify that

$$\begin{aligned} & L(\hat{g}_m|A_1) = \mathbb{P}(\hat{g}_m(X) \neq Y|A_1) = \mathbb{P}(g^-(X) \neq Y|A_1) = 0 \\ \implies \quad & \mathbb{P}(L(\hat{g}_m) > \varepsilon|A_1) = \mathbb{P}(0 > \varepsilon) = 0 < \delta \end{aligned} \tag{40}$$

for all $(\varepsilon, \delta) \in (0,1)^2$. Thus, part (ii) is trivially true when the perfect classifier is g^-.

Case 2: The perfect classifier is g_z for some $z \in S$ and the distribution of X is s.t. $\mathbb{P}(X = z) \leq \varepsilon$. Let A_{21} = the event that $z \in \{X_1, ..., X_n\}$. Then

$$\mathbb{P}(L(\hat{g}_m) > \varepsilon|A_2) = \mathbb{P}(L(\hat{g}_m) > \varepsilon|A_2, A_{21})\mathbb{P}(A_{21}|A_2) + \mathbb{P}(L(\hat{g}_m) > \varepsilon|A_2, A_{21}^c)\mathbb{P}(A_{21}^c|A_2) \tag{41}$$

(i) Consider the case where $z \in \{X_1, ..., X_n\}$. Then we know $z = X_i$ for some $i \in \{1, ..., n\}$ and $g_z(X_i) = Y_i = 1$ since g_z is the perfect classifier and we assume all instances from the training data are labeled correctly. Since $|\{i \in \{1, ..., n\}|Y_i = 1\}| \in \{0, 1\}$ by assumption, and we know $Y_i = 1$, we know we have $|\{i \in \{1, ..., n\}|Y_i = 1\}| = 1$. By definition, the output of our modified ERM algorithm will be $\hat{g}_m := g_{X_i} = g_z$. Thus, if g_z is the perfect classifier, $z \in \{X_1, ..., X_n\}$, and $\mathbb{P}(X = z) \leq \varepsilon$, our modified ERM algorithm will always output the perfect classifier. We can easily verify that

$$\begin{aligned} & L(\hat{g}_m|A_2, A_{21}) = \mathbb{P}(\hat{g}_m(X) \neq Y|A_2, A_{21}) = \mathbb{P}(g_z(x) \neq Y|A_2, A_{21}) = 0 \\ \implies \quad & \mathbb{P}(L(\hat{g}_m) > \varepsilon|A_2, A_{21}) = \mathbb{P}(0 > \varepsilon) = 0 < \delta \end{aligned} \tag{42}$$

for all $(\varepsilon, \delta) \in (0,1)^2$. Thus, part (ii) of the proof of PAC learnability is trivially true when the perfect classifier is g_z, $z \in \{X_1, ..., X_n)\}$, and $\mathbb{P}(X = z) \leq \varepsilon$.

(ii) Consider now $z \notin \{X_1, ..., X_n\}$. Then, assuming all instances from the training data are properly labeled, we have $|\{i \in \{1, ..., n\}|Y_i = 1\}| = 0$. This follows from the fact that $\mathbb{P}(g_z(X) = Y) = 1$ and $g_z(X_i) = -1$ for all $i \in \{1, ..., n\}$. By definition, our modified ERM algorithm will output $\hat{g}_m := g^-$. Note that $g_z(X) = g^-(X)$ for all $X \neq z$. This implies that

$$L(g^-|A_2, A_{21}^c) = \mathbb{P}(g^-(X) \neq Y|A_2, A_{21}^c) = \mathbb{P}(X = z|A_2, A_{21}^c)$$

We are given that $\mathbb{P}(X = z) \leq \varepsilon$, so we find

$$L(g^-|A_2, A_{21}^c) \leq \varepsilon$$

for all $\varepsilon \in (0, 1)$. Thus,

$$\mathbb{P}(L(g^-) > \varepsilon|A_2, A_{21}^c) = 1 - \mathbb{P}(L(g^-) \leq \varepsilon|A_2, A_{21}^c) = 1 - 1 = 0$$

We can easily verify that

$$\mathbb{P}(L(\hat{g}_m) > \varepsilon|A_2, A_{21}^c) = \mathbb{P}(L(g^-) > \varepsilon|A_2, A_{21}^c) = 0 < \delta \tag{43}$$

for all $(\varepsilon, \delta) \in (0,1)^2$. Thus, part (ii) of the proof of PAC learnability is trivially true when the perfect classifier is g_z, $z \notin \{X_1, ..., X_n\}$, and $\mathbb{P}(X = z) \leq \varepsilon$.

Parts (i) and (ii) of Case 2 combine to prove part (ii) of the proof of PAC learnability is trivially true when the perfect classifier is g_z and $\mathbb{P}(X = z) \leq \varepsilon$. Plugging (42) and (43) into (41) yields

$$\mathbb{P}(L(\hat{g}_m) > \varepsilon|A_2) = 0 + 0 = 0 \tag{44}$$

Case 3: The perfect classifier is g_z for some $z \in S$ and the distribution of X is s.t. $\mathbb{P}(X = z) > \varepsilon$. Let A_{31} = the event that $z \in \{X_1, ..., X_n\}$. Note that

$$\begin{aligned} \mathbb{P}(L(\hat{g}_m) > \varepsilon|A_3) = \quad & \mathbb{P}(L(\hat{g}_m) > \varepsilon|A_3, A_{31})\mathbb{P}(A_{31}|A_3) \\ + \quad & \mathbb{P}(L(\hat{g}_m) > \varepsilon|A_3, A_{31}^c)\mathbb{P}(A_{31}^c|A_3) \end{aligned} \tag{45}$$

(i) Consider $z \in \{X_1, .., X_n\}$. From Case 2, part (i), we know $|\{i \in \{1, ..., n\} | Y_i = 1\}| = 1$ and $z = X_i$ for some $i \in \{1, ..., n\}$. By definition, our modified ERM algorithm will output $\hat{g}_m := g_{X_i} = g_z$. Thus, if g_z is the perfect classifier, $z \in \{1, ..., n\}$, and $\mathbb{P}(X = z) > \varepsilon$, our modified ERM algorithm will always output the perfect classifier. We can easily verify that

$$\begin{aligned} & L(\hat{g}_m | A_3, A_{31}) = \mathbb{P}(\hat{g}_m(X) \neq Y | A_3, A_{31}) = \mathbb{P}(g_z(x) \neq Y | A_3, A_{31}) = 0 \\ \implies \quad & \mathbb{P}(L(\hat{g}_m) > \varepsilon | A_3, A_{31}) = \mathbb{P}(0 > \varepsilon) = 0 < \delta \qquad (46) \end{aligned}$$

for all $(\varepsilon, \delta) \in (0, 1)^2$. Thus, part (ii) of the proof of PAC learnability is trivially true when the perfect classifier is g_z, $z \in \{X_1, .., X_n\}$, and $\mathbb{P}(X = z) > \varepsilon$.

(ii) Consider $z \in \{X_1, ..., X_n\}$. From Case 2, part (ii), we know $|\{i \in \{1, ..., n\} | Y_i = 1\}| = 0$. By definition, our modified ERM algorithm will output $\hat{g}_m := g^-$. Note that $g_z(X) = g^-(X)$ for all $X \neq z$. This implies that

$$L(g^- | A_3, A_{31}^c) = \mathbb{P}(g^-(X) \neq Y | A_3, A_{31}^c) = \mathbb{P}(X = z | A_3, A_{31}^c)$$

We are given that $\mathbb{P}(X = z) > \varepsilon$, so we find

$$L(g^- | A_3, A_{31}^c) > \varepsilon$$

for all $\varepsilon \in (0, 1)$. Thus, we know

$$\mathbb{P}(L(\hat{g}_m) > \varepsilon | A_3, A_{31}^c) = \mathbb{P}(L(g^-) > \varepsilon | A_3, A_{31}^c) = 1 \qquad (47)$$

Plugging (47) and (46) into (45) yields

$$\mathbb{P}(L(\hat{g}_m) > \varepsilon | A_3) = 0 + \mathbb{P}(A_{31}^c | A_3) \qquad (48)$$

Since we are given (via A_3) that $\mathbb{P}(X = z) > \varepsilon$, we know $\mathbb{P}(X \neq z | A_3) = 1 - \mathbb{P}(X = z | A_3) < 1 - \varepsilon$. Since $X_1, ..., X_n$ are independent and identically distributed, we know

$$\mathbb{P}(A_{31}^c | A_3) = \mathbb{P}(z \notin \{X_1, ..., X_n\} | A_3) = \mathbb{P}(X_1 \neq z, ..., X_n \neq z | A_3) = \mathbb{P}(X_1 \neq z | A_3)^n < (1 - \varepsilon)^n \qquad (49)$$

Plugging (49) into (48) yields

$$\mathbb{P}(L(\hat{g}_m) > \varepsilon | A_3) < (1 - \varepsilon)^n \qquad (50)$$

Plugging (50), (44), and (40) into (39) yields

$$\mathbb{P}(L(\hat{g}_m) > \varepsilon) = \mathbb{P}(L(\hat{g}_m) > \varepsilon | A_3) \mathbb{P}(A_3) \qquad (51)$$

Applying the fact that $\mathbb{P}(A_3) \leq 1$ by the axioms of probability, along with the bound from (50), we have

$$\mathbb{P}(L(\hat{g}_m) > \varepsilon) \leq \mathbb{P}(L(\hat{g}_m) > \varepsilon | A_3) < (1 - \varepsilon)^n$$

Note that

$$(1 - \varepsilon)^n < \delta \iff n ln(1 - \varepsilon) < ln(\delta) \iff n > \frac{ln(\delta)}{ln(1 - \varepsilon)}$$

with the last inequality following from the fact that $ln(1 - \varepsilon) < 0$ for all $\varepsilon \in (0, 1)$. Thus, to guarantee that

$$P(L(\hat{g}_m) > \varepsilon) < (1 - \varepsilon)^n < \delta$$

for any $(\varepsilon, \delta) \in (0, 1)^2$, we just need to ensure

$$n > \frac{ln(\delta)}{ln(1 - \varepsilon)}$$

This completes the proof that G is PAC learnable under the modified ERM algorithm with output $\hat{g}_m$ and sample complexity

$$n > n(\varepsilon, \delta) = \frac{ln(\delta)}{ln(1 - \varepsilon)}$$

for all $(\varepsilon, \delta) \in (0, 1)^2$ where $n \in \mathbb{N}$.

4.

The goal of this exercise is to prove Hoeffding's inequality. Do as many parts as you can, and feel free to skip the steps you are uncertain about (you may just assume that they are true and move on to the next steps). Let $X_1, ..., X_n$ be i.i.d. random variables such that $\mathbb{E}[X_1] = 0$ and $a \leq X_1 \leq b$ with probability 1.

(a) Demonstrate that $a \leq 0$ and $b \geq 0$.

(b) For the rest of the problem, we will assume that b-a =1 (if not, replace X_j by $\frac{X_j}{b-a}$ for all j). Let $\lambda > 0$ be any positive number. Then the function $f(x) = e^{\lambda x}$ is convex, meaning that

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$$

for any x,y and $\alpha \in [0,1]$. (You don't need to prove it). Write

$$X = \underbrace{(b-X)}_{=\alpha} \cdot a + \underbrace{(X-a)}_{=1-\alpha} \cdot b$$

and demonstrate that

$$\mathbb{E}[e^{\lambda X}] \leq e^{-\lambda(-a)}b + e^{\lambda b}(-a) = e^{-\lambda(1-b)}b + e^{\lambda b}(1-b) = e^{\lambda b}(1-b+be^{-\lambda})$$

(c) Let

$$F(\lambda) = ln(e^{\lambda b}(1-b+be^{-\lambda}))$$

(log with base e). Show that $F'(0) = 0$ and that $F"(0) \leq \frac{1}{4}$ for any $b \in [0,1]$. Conclude from this and the Taylor's expansion that

$$F(\lambda) \leq \frac{\lambda^2}{8}$$

(d) So far, we've shown that $\mathbb{E}[e^{\lambda x}] \leq e^{\frac{\lambda^2}{8}}$. Now recall Markov's inequality: For any positive random variable Z, $\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}[Z]}{t}$. Use it to show that

$$\mathbb{P}(X_1 + \cdots + X_n \geq t) = \mathbb{P}(e^{\lambda(X_1+...+X_n)} \geq e^{\lambda t}) \leq \mathbb{E}[e^{\lambda(X_1+...+X_n)}]e^{-\lambda t}$$

Note that this bound is valid for any $\lambda > 0$.

(e) Show, using independence, that $\mathbb{E}[e^{\lambda(X_1+\cdots+X_n)}] \leq e^{\frac{n\lambda^2}{8}}$.

(f) Finally, combine parts (d) and (e) and choose the value of λ that minimizes $\mathbb{E}[e^{\lambda(X_1+\cdots+X_n)}]e^{-\lambda t}$, and write the resulting bound for

$$\mathbb{P}(X_1 + \cdots + X_n \geq t)$$

Congratulations, you have just proven Hoeffding's inequality.

Solution.

(a) First, we will prove $a \leq 0$. Assume to the contrary that $a > 0$. Then

$$\mathbb{P}(a \leq X \leq b) = 1 \implies \mathbb{P}(X < a) = 0 \implies \mathbb{P}(X < 0 < a) = 0$$

since $(X < 0) \subseteq (X < a) \implies \mathbb{P}(X < 0) \leq \mathbb{P}(X < a) = 0$ and $\mathbb{P}(X < 0) \geq 0$ by the axioms of probability. Let D^+ be the set of values which X can take with nonzero probability under the assumption $a > 0$. Then $\forall d \in D^+$, we know $d > 0$. If X is discrete, we have

$$\mathbb{E}[X] = \sum_{d \in D^+} d\mathbb{P}(X = d) > \sum_{d \in D^+} 0 = 0$$

since $\mathbb{P}(X = d) \geq 0$ by the axioms of probability, but we are given

$$\mathbb{E}[X] = 0 \not> 0$$

so we have derived a contradiction. This completes the proof that $a \leq 0$ for discrete X. If X is a continuous random variable, we know $D^+ = [a, b]$, so

$$E[X] = \int_a^b x f_X(x) dx > \int_a^b 0 dx = 0$$

since $x > 0$ for all $0 < a \leq x \leq b$ and $f_X(x) \geq 0$ by definition of the probability density function. One again, since we are given

$$\mathbb{E}[X] = 0 \not> 0$$

we have derived a contradiction. This completes the proof that $a \leq 0$ for continuous X, and combines with the previous proof for discrete X to prove that $a \leq 0$ for all random variables X.
Now, we will prove $b \geq 0$. Similarly, assume to the contrary that $b < 0$. Then

$$\mathbb{P}(a \leq X \leq b) = 1 \implies \mathbb{P}(X > b) = 0 \implies \mathbb{P}(X > 0 > b) = 0$$

since $(X > 0) \subseteq (X > b) \implies \mathbb{P}(X > 0) \leq \mathbb{P}(X > b) = 0$ and $\mathbb{P}(X > 0) \geq 0$ by the axioms of probability. Now, define D^- to be the set of values which X can take with nonzero probability under the assumption $b < 0$. Then $\forall d \in D^-$, we know $d < 0$. If X is discrete, we have

$$\mathbb{E}[X] = \sum_{d \in D^-} x\mathbb{P}(X = x) < \sum_{d \in D^-} 0 = 0$$

since $\mathbb{P}(X = x) \geq 0$ by the axioms of probability. Since we are given

$$\mathbb{E}[X] = 0 \not> 0$$

we have derived a contradiction. This completes the proof that $b \geq 0$ for all discrete X. If X is continuous, then $D^- = [a, b]$, so we have

$$\mathbb{E}[X] = \int_a^b x f_X(X) dx < \int_a^b 0 dx = 0$$

since $x < 0$ for all $a \leq x \leq b < 0$ and $f_X(x) \geq 0$ by the definition of the probability mass function. Again, since we are given

$$\mathbb{E}[X] = 0 \not> 0$$

we have derived a contradiction, which completes the proof that $b \geq 0$ for all continuous X, and combines with the previous proof to complete the proof that $b \geq 0$ for all random variables X.

(b) Note that $a \leq X \leq b$ with probability 1 combines with our assumption that $b - a = 1$ to imply

$$b - b = 0 \leq b - X \leq b - a = 1 \qquad a - a = 0 \leq X - a \leq b - a = 1$$

Thus, we can write

$$f(x) = f(\underbrace{(b - X)}_{=\alpha} \cdot a + \underbrace{(X - a)}_{=1-\alpha} \cdot b)$$

and apply the fact that $f(x)$ is a convex function and

$$X = \alpha x + (1 - \alpha) y$$

with $\alpha = b - X$, $x = a$, and $y = b$ to find

$$\begin{aligned} f(X) = \quad & e^{\lambda X} = f(\alpha x + (1 - \alpha) y) = f((b - X)a + (X - a)b) \\ \leq \quad & \alpha f(x) + (1 - \alpha) f(y) = (b - X)e^{\lambda a} + (X - a)e^{\lambda b} = be^{\lambda a} - Xe^{\lambda a} + Xe^{\lambda b} - ae^{\lambda b} \end{aligned} \tag{52}$$

Taking the expectation of (52) yields

$$\mathbb{E}[f(X)] = \mathbb{E}[e^{\lambda X}] \leq \mathbb{E}[be^{\lambda a} - Xe^{\lambda a} + Xe^{\lambda b} - ae^{\lambda b}] \tag{53}$$

Applying Linearity of Expectation and the fact that $\mathbb{E}[X] = 0$ to (53) yields

$$\mathbb{E}[e^{\lambda X}] \leq be^{\lambda a} - e^{\lambda a}\mathbb{E}[X] + e^{\lambda b}\mathbb{E}[X] - ae^{\lambda b} = be^{\lambda a} - ae^{\lambda b} = e^{\lambda a}b + e^{\lambda b}(-a) \tag{54}$$

Noting that $\lambda a = -\lambda(-a)$, $-a = 1 - b$, and simplifying (54) yields

$$\mathbb{E}[e^{\lambda X}] \leq e^{-\lambda(-a)}b + e^{\lambda b}(-a) = e^{-\lambda(1-b)}b + e^{\lambda b}(1-b) = e^{\lambda b}(1 - b + be^{-\lambda})$$

which completes the proof for part (b).

(c) First, we will show that $F'(0) = 0$. First, note that, since $ln(ab) = ln(a) + ln(b)$, for nonnegative numbers $a, b \in \mathbb{R}$, and $e^{\lambda b} \geq 0$ for all $b \in [0,1]$, λ and $(1 - b + be^{-\lambda}) \geq 0$ for all $b \in [0,1]$, λ, we have

$$F(\lambda) = ln(e^{\lambda b}(1 - b + be^{-\lambda}) = ln(e^{\lambda b}) + ln(1 + b + be^{-\lambda}) = \lambda b + ln(1 + b + be^{-\lambda}) \tag{55}$$

Differentiating (55) with respect to λ yields

$$F'(\lambda) = \frac{d}{d\lambda}(\lambda b) + \frac{d}{d\lambda}(ln(1 - b + be^{-\lambda})) = b + \frac{-be^{-\lambda}}{1 - b + be^{-\lambda}} \tag{56}$$

Evaluating (56) at $\lambda = 0$ yields

$$F'(0) = b + \frac{-b \cdot 1}{1 - b + b \cdot 1} = b + \frac{-b}{1 - b + b} = b + \frac{-b}{1} = b - b = 0$$

which completes the proof that $F'(0) = 0$.
Now, we will show that $F"(0) \leq \frac{1}{4}$ for all $b \in [0,1]$. Differentiating (56) with respect to λ yields

$$F"(\lambda) = \frac{d}{d\lambda}(b + \frac{-be^{-\lambda}}{1 - b + be^{-\lambda}}) = \frac{d}{d\lambda}(\frac{-be^{-\lambda}}{1 - b + be^{-\lambda}}) = \frac{be^{-\lambda}(1 - b + be^{-\lambda}) - (-be^{-\lambda})(-be^{-\lambda})}{(1 - b + be^{-\lambda})^2} \tag{57}$$

Simplifying the numerator of (57) yields

$$be^{-\lambda}(1-b+be^{-\lambda}) - (-be^{-\lambda})(-be^{-\lambda}) = be^{-\lambda} - b^2e^{-\lambda} + b^2e^{-2\lambda} - b^2e^{-2\lambda} = be^{-\lambda} - b^2e^{-\lambda} = (1-b)be^{-\lambda} \tag{58}$$

Plugging (58) into (57) yields

$$F"(\lambda) = \frac{(1-b)be^{-\lambda}}{(1 - b + be^{-\lambda})^2} \tag{59}$$

Evaluating (59) at $\lambda = 0$ yields

$$F"(0) = \frac{(1-b)b \cdot 1}{(1 - b + b \cdot 1)^2} = \frac{(1-b)b}{1 - b + b} = (1-b)b \tag{60}$$

Note that

$$(1-b)b = -b^2 + b$$

is a quadratic in b opening down with global maximum found when $\frac{d}{db}(-b^2 + b) = 0$. We can easily compute

$$\frac{d}{db}(-b^2 + b) = -2b + 1 = 0 \implies -2b = -1 \implies b = \frac{1}{2}$$

so we know

$$F"(0) = -b^2 + b \leq -(\frac{1}{2})^2 + \frac{1}{2} = -\frac{1}{4} + \frac{1}{2} = \frac{1}{4} \tag{61}$$

This completes the proof that $F"(0) \leq \frac{1}{4}$ for all $b \in [0,1]$. Applying the Taylor expansion of $F(\lambda)$ at $a = 0$ yields

$$F(\lambda) = F(0) + \lambda F'(0) + \frac{\lambda^2}{2} F"(c) \qquad (62)$$

where $c \in \mathbb{R}$ s.t. $0 \leq c \leq \lambda$. Thus, to get an upper bound for $F(\lambda)$, we need to find an upper bound for $F"(\lambda)$ for all $\lambda \in \mathbb{R}$. By the AMGM inequality, we know

$$\frac{x+y}{2} \geq \sqrt{xy} \qquad (63)$$

for all $x, y \in \mathbb{R}$. Setting $x = (1-b)$, $y = be^{-\lambda}$ and applying (63), we find

$$\begin{aligned} & \sqrt{(1-b)be^{-\lambda}} \leq \frac{(1-b)+be^{-\lambda}}{2} \\ \implies \quad & (1-b)be^{-\lambda} \leq \frac{(1-b+be^{-\lambda})^2}{4} \\ \implies \quad & F"(\lambda) = \frac{(1-b)be^{-\lambda}}{(1-b+be^{-\lambda})^2} \leq \frac{1}{4} \qquad (64) \end{aligned}$$

for all $\lambda \in \mathbb{R}$. Noting that $F(0) = ln(1 \cdot (1)) = 0$ and plugging (64) and $F'(0) = 0$ into (62) yields

$$F(\lambda) = F(0) + \lambda F'(0) + \frac{\lambda^2}{2} F"(c) = \frac{\lambda^2}{2} F"(c) \leq \frac{\lambda^2}{2}\frac{1}{4} = \frac{\lambda^2}{8} \qquad (65)$$

This completes the proof that $F(\lambda) \leq \frac{\lambda^2}{8}$ for all $\lambda \in \mathbb{R}$.

(d) Note that

$$X_1 + \cdots + X_n \geq t \iff e^{\lambda(X_1+\cdots+X_n)} \geq e^{\lambda t}$$

so we can write

$$\mathbb{P}(X_1 + \cdots + X_n \geq t) = \mathbb{P}(e^{\lambda(X_1+\cdots+X_n)} \geq e^{\lambda t}) \qquad (66)$$

Since $e^x \geq 0$ for all $x \in \mathbb{R}$, we can apply Markov's inequality with $t = e^{\lambda t}$ to (66) to find

$$\mathbb{P}(X_1 + \cdots + X_n \geq t) \leq \frac{\mathbb{E}[e^{\lambda(X_1+\cdots+X_n)}]}{e^{\lambda t}} = \mathbb{E}[e^{\lambda(X_1+\cdots+X_n)}]e^{-\lambda t} \qquad (67)$$

This completes the proof that

$$\mathbb{P}(X_1 + \cdots + X_n \geq t) = \mathbb{P}(e^{\lambda(X_1+\cdots+X_n)} \geq e^{\lambda t}) \leq \mathbb{E}[e^{\lambda(X_1+\cdots+X_n)}]e^{-\lambda t}$$

for all $\lambda > 0$.

(e) First, note that

$$e^{\lambda(X_1+\cdots+X_n)} = e^{\lambda X_1+\cdots+\lambda X_n} = e^{\lambda X_1} \cdots e^{\lambda X_n} \qquad (68)$$

Since $X_1, ..., X_n$ are independent, and $e^{\lambda X_i}$ depends only on X_i and is independent of X_j for $i \neq j \in \{1, ..., n\}$, we know $e^{\lambda X_1}, ..., e^{\lambda X_n}$ are independent. For all independent random variables $X_1, ..., X_n$, we know

$$\mathbb{E}[X_1 \cdots X_n] = \mathbb{E}[X_1] \cdots \mathbb{E}[X_n] \qquad (69)$$

Applying (69) to (68) yields

$$\mathbb{E}[e^{\lambda(X_1+\cdots+X_n)}] = \mathbb{E}[e^{\lambda X_1}] \cdots \mathbb{E}[e^{\lambda X_n}] \qquad (70)$$

Since $X_1, ..., X_n$ are identically distributed, we know

$$\mathbb{E}[e^{\lambda X_1}] = \cdots = \mathbb{E}[e^{\lambda X_n}] \qquad (71)$$

Plugging (71) into (70) yields

$$\mathbb{E}[e^{\lambda(X_1+\cdots+X_n)}] = (\mathbb{E}[e^{\lambda X_1}])^n \tag{72}$$

We know from part (c) that $\mathbb{E}[e^{\lambda X}] \leq e^{F(\lambda)} \leq e^{\frac{\lambda^2}{8}}$. Plugging this into (72) yields

$$\mathbb{E}[e^{\lambda(X_1+\cdots+X_n)}] \leq (e^{\frac{\lambda^2}{8}})^n = e^{\frac{n\lambda^2}{8}}$$

which completes the proof for (e).

(f) From part (d), we know

$$\mathbb{P}(X_1 + \cdots + X_n \geq t) \leq \mathbb{E}[e^{\lambda(X_1+\cdots+X_n)}]e^{-\lambda t} \tag{73}$$

for all $\lambda > 0$, which means (73) holds for

$$\lambda^* := \lambda \in \mathbb{R} \text{ s.t. } \mathbb{E}[e^{\lambda(X_1+\cdots+X_n)}]e^{-\lambda t} \text{ is minimized over all } \lambda > 0.$$

Plugging λ^* for λ in (73) yields

$$\mathbb{P}(X_1 + \cdots + X_n \geq t) \leq \mathbb{E}[e^{\lambda^*(X_1+\cdots+X_n)}]e^{-\lambda^* t} \tag{74}$$

By the definition of λ^*, we know

$$\mathbb{E}[e^{\lambda^*(X_1+\cdots+X_n)}]e^{-\lambda^* t} \leq \mathbb{E}[e^{\lambda(X_1+\cdots+X_n)}]e^{-\lambda t} \tag{75}$$

for all $\lambda > 0$. Thus, we can let $\lambda = \frac{4t}{n} > 0$ for all $t, n > 0$ and apply (75) and the result from part (e) to (74) to find

$$\mathbb{P}(X_1 + \cdots + X_n \geq t) \leq \mathbb{E}[e^{\lambda(X_1+\cdots+X_n)}]e^{-\lambda t} \leq e^{\frac{n\lambda^2}{8}-\lambda t} = e^{\frac{16nt^2}{8n^2}-\frac{4t^2}{n}} = e^{\frac{2t^2}{n}-\frac{4t^2}{n}} = e^{\frac{-2t^2}{n}} \tag{76}$$

By symmetry, we have

$$\mathbb{P}(-(X_1 + \cdots + X_n) \geq t) = \mathbb{P}(X_1 + \cdots + X_n) \leq -t) \leq e^{\frac{-2t^2}{n}} \tag{77}$$

so we know

$$\mathbb{P}(|X_1 + \cdots + X_n| \geq t) = \mathbb{P}(X_1 + \cdots + X_n) \leq -t) + \mathbb{P}(X_1 + \cdots + X_n) \geq t) \leq 2e^{\frac{-2t^2}{n}} \tag{78}$$

Note that

$$\mathbb{P}(|X_1 + \cdots + X_n| \geq t) = \mathbb{P}(|\frac{X_1 + \cdots + X_n}{n}| \geq \frac{t}{n}) \tag{79}$$

Hoeffding's inequality follows from letting $t = nt_1$ to find

$$\mathbb{P}(|\frac{X_1 + \cdots + X_n}{n}| \geq t_1) \leq 2e^{\frac{-2(nt_1)^2}{n}} = 2e^{\frac{-2n^2t_1^2}{n}} = 2e^{-2nt_1^2} \tag{80}$$

Noting that

$$2e^{-2nt_1^2} = 2e^{\frac{-2nt_1^2}{(b-a)^2}}$$

since we assume $b - a = 1$ completes the proof of Hoeffding's inequality.

Assignment 3

MATH 447: Homework 3

Read chapter 6 of the textbook "Understanding Machine Learning." Then do the following problems:

1. Let G be a base class such that $|G| < \infty$, that is, G has finitely many elements. Does G have finite Vapnik-Chervonenkis (VC) dimension? If so, give an upper bound for VC(G). If not, provide an example.

   Solution.
   Claim: G does have finite VC dimension, and $VC(G) \leq \log_2(|G|) < \infty$.
   Proof. For any collection of instances $C = \{x_1, ..., x_n\}$ of size $|C| = n$, G shatters $C \iff$

   $$|G_C| = |\{(g(x_1), ..., g(x_n)) : g \in G\}| = 2^n$$

   That is, G only shatters C if for all $C_1 \subseteq C$, $\exists g \in G$ such that $g(x) = +1$ for all $c \in C_1$ and $g(x) = -1$ for all $c \in C \setminus C_1$. For any fixed C, the subsets $A \subseteq C$ and $C \setminus A$ such that $g(x) = +1$ for all $c \in A$ and $g(x) = -1$ for all $c \in C \setminus A$ are fixed. Thus, each $g \in G$ will correspond to exactly one sequence $(g(x_1), ..., g(x_n))$ (A consists of the x_i's mapped to $+1$ by g, while $C \setminus A$ consists of all other $x_i \in C$). There are exactly 2^n subsets $A \subseteq C$, $|C| = n$, so there are 2^n possible sequences $(s_1, ..., s_n) \in \{-1, +1\}^n$, and each g corresponds to exactly one such sequence, so to produce all 2^n sequences using only $g \in G$ (and thus to have $|G_C| = 2^n$), we must have at least 2^n unique classifiers $g \in G$. That is, we must have

   $$|G| \geq 2^n$$

   By the definition of VC dimension,

   $$VC(G) := \max\{n \in \mathbb{N} : \tau_G(n) = 2^n\}$$

   and by the definition of the growth function $\tau_G(n)$, we know

   $$\tau_G(n) = 2^n \iff \exists \text{ a collection of instances } C \text{ s.t. } |C| = n \text{ and } |G_C| = 2^n$$

   Consider the maximum size collection C such that $|C| = n$, $|G_C| = 2^n$. By definition, the VC dimension of G is the size of this collection C, so $n = VC(G)$. We just showed that for any such C where $|C| = n$ and $|G_C| = 2^n$, $|G| \geq 2^n$. Thus,

   $$|G| \geq 2^{VC(G)}$$

   Taking $\log_2$ of both sides yields

   $$\log_2(|G|) \geq \log_2(2^{VC(G)}) = VC(G)$$

   We are given $|G| < \infty$, so we know $\log_2(|G|) < \infty$, so we can provide a finite upper bound for the VC dimension of G with

   $$VC(G) \leq \log_2(|G|) < \infty$$

   This completes the proof that if $|G| < \infty$, then G has finite VC dimension, upper bounded by $\log_2(|G|) < \infty$.

2. (a) Let S be an infinite discrete set, and consider the family of binary classifiers G that consists of all functions $g_z(x) = \begin{cases} 1, & x = z, \\ -1, & x \neq z, \end{cases}$ indexed by $z \in S$, and also includes the classifier $g^-(x)$ that is identically equal to -1. In the previous homework, you showed that this class is PAC learnable. Find the VC dimension of the class G using the definition of VC dimension.

(b) Let $k > 1$ be a fixed integer, and consider the set G of all binary classifiers g such that g takes value $+1$ in exactly k points. What is the VC dimension of G?

Solution.

(a) *Claim:* The VC dimension of G is $VC(G) = 1$.
Proof. By definition of the VC dimension,

$$VC(G) := \max\{n \in \mathbb{N} : \tau_G(n) = 2^n\}$$

where τ_G is the growth function of the family G. Thus, to prove that $VC(G) = 1$, we just need to:
i) Find a set of instances C such that $|C| = 1$ and $|G_C| = 2^1 = 2$ (find a set of size 1 such that G shatters C).
ii) Prove that for all sets C such that $|C| = 1 + 1 = 2$, $|G_C| < 2^2 = 4$ (show all sets of size 1+1 are *not* shattered by G.
First, consider a set $C = \{x\}$ such that $x \in S$. Then $g_x \in G$ and $(g_x(x_1), ..., g_x(x_n)) = g_x(x) = +1$. Also, $g^- \in G$ and $(g^-(x_1), ..., g^-(x_n)) = g^-(x) = -1$. Thus, we have

$$G_C = \{(g(x_1), ..., g(x_n))|g \in G\} = \{(+1), (-1)\}$$

so

$$|G_C| = 2 = 2^1$$

so G shatters C. This completes the first part of the proof.
Now, consider an arbitrary set $C = \{x_1, x_2\}$. Since

$$\begin{aligned} & |G_C| = 2^2 = 4 \\ \iff \quad & G_C = \{(g(x_1), ..., g(x_n))|g \in G\} = \{(+1, +1), (-1, -1), (+1, -1), (-1, +1)\} \end{aligned}$$

it suffices to show that $(g(x_1), ..., g(x_n)) = (g(x_1), g(x_2)) \neq (+1, +1)$ for all $g \in G$.
If $x_1 \notin S$, then $g_{x_1} \notin G$, so for all $g \in G$, $g(x_1) = -1 \neq +1$, so $(g(x_1), g(x_2)) \neq (+1, +1)$ for all $g \in G$.
If $x_2 \notin S$, then $g_{x_2} \notin G$, so for all $g \in G$, $g(x_1) = -1 \neq +1$, so $(g(x_1), g(x_2)) \neq (+1, +1)$ for all $g \in G$.
If both $x_1, x_2 \in S$, then $g_{x_1}, g_{x_2} \in G$, but $g_{x_1}(x_2) := -1 \neq +1$ and $g_{x_2}(x_1) = -1 \neq +1$, and for all other $g \in G$, $g(x_1) = g(x_2) = -1 \neq +1$, so for all $g \in G$, $(g(x_1), g(x_2)) \neq (+1, +1)$.
Thus, regardless of the composition of S, we know $(g(x_1), g(x_2)) \neq (+1, +1)$ for all $g \in G$, so we know $|G_C| < 4$ for all $C = \{x_1, x_2\}$ (i.e. all C such that $|C| = 2$). This completes the proof that G cannot shatter any collection C of size $|C| = 2$.
We already found a set C of size $|C| = 1$ such that G shatters C, so this completes the proof that the VC dimension of G is

$$VC(G) = 1$$

(b) *Claim:* The VC dimension of G is $VC(G) = k$.
Proof. Once again, by the definition of the VC dimension, to prove

$$VC(G) := \max\{n \in \mathbb{N} : \tau_G(n) = 2^n\} = k$$

we must:
i) Find a collection $C = \{x_1, ..., x_k\}$ of size k such that G shatters C (i.e. $|G_C = 2^k$)
ii) Show that all collections $C = \{x_1, ..., x_{k+1}\}$ of size $k + 1$ are *not* shattered by G (i.e. $|G_C| < 2^{k+1}$).
First, we will show part (i). Consider any arbitrary collection of points $C = \{x_1, ..., x_k\}$ of size k. Pick another arbitrary collection of points $D = \{y_1, ..., y_k\}$ of size k such that

$$C \cap D = \emptyset$$

Note that, since G is defined to consist of *all* binary classifiers g taking value $+1$ in exactly k points, we know $\exists g_C \in G$ such that

$$g_C(x) = \begin{cases} +1 \text{ for all } x \in C \\ -1 \text{ otherwise} \end{cases}$$

and $\exists g_D \in G$ such that

$$g_D(x) = \begin{cases} +1 \text{ for all } x \in D \\ -1 \text{ otherwise} \end{cases}$$

Consider any arbitrary subset $A \subseteq C$ where $|A| = i$ for some $i \in \{0, ..., k\}$ (i.e. all subsets of C including $\emptyset$ and C itself). Arbitrary pick $B \subseteq D$ such that $|B| = k - i$. Then $|A \cup B| = k$ because $A \subseteq C$, $B \subseteq D$, and $C \cap D = \emptyset$. By the definition of G, we know $\exists g_{A \cup B} \in G$ such that

$$g_{A \cup B}(x) = \begin{cases} +1 \text{ for all } x \in A \cup B \\ -1 \text{ otherwise} \end{cases}$$

Since $B \cap C = \emptyset$, for all $x \in C$, we have

$$g_{A \cup B}(x) = \begin{cases} +1 \text{ for all } x \in A \\ -1 \text{ for all } x \in C \setminus A \end{cases}$$

Since this holds for arbitrary $A \subseteq C$ of size $|A| = i \in \{0, ..., k\}$, we know it holds for all 2^k subsets $A \subseteq C$. Since

$$(g_{A_1 \cup B_1}(x_1), ..., g_{A_1 \cup B_1}(x_k)) \neq (g_{A_2 \cup B_2}(x_1), ..., g_{A_2 \cup B_2}(x_k))$$

for all $A_1 \neq A_2$, and there are 2^k distinct subsets $A \subseteq C$, we know we can produce 2^k distinct sequences $(g(x_1), ..., g(x_k))$ using only $g \in G$. Thus, we know

$$|G_C| := |\{(g(x_1), ..., g(x_k) | g \in G\}| = 2^k$$

which completes the proof that G shatters all C of size $|C| = k$.
Now, consider an arbitrary C of size $|C| = k + 1$. Since there are exactly 2^{k+1} sequences of signs of the form $(s_1, ..., s_{k+1}) \in \{-1, +1\}^{k+1}$, to prove

$$|G_C| := |\{(g(x_1), ..., g(x_{k+1})) | g \in G\}| < 2^{k+1}$$

it suffices to show that $(g(x_1), ..., g(x_{k+1})) \neq (+1, ..., +1) = (s_1, ..., s_{k+1})$ for all $g \in G$ (i.e. no $g \in G$ can map all $k+1$ $x \in C$ to $+1$). By definition of G, for all $g \in G$, we know $g(x) = +1$ only for all $x \in A$ where A is a set of size k. Thus, regardless of the specific points in C, for all $g \in G$, we have

$$|\{g(x) | g(x) = +1 \quad x \in C\}| = |\{s | s \in (g(x_1), ..., g(x_{k+1})) \quad s = +1\}| =\leq k$$

However, by definition, in the sequence of $k + 1$ $+1$'s $(s_1, ..., s_{k+1}) = (+1, ..., +1)$,

$$|\{s | s \in (s_1, ..., s_{k+1}) \quad s = +1\}| = k + 1$$

Thus,

$$|\{g(x) | g(x) = +1 \quad x \in C\}| \neq |\{s | s \in (s_1, ..., s_{k+1}) \quad s = +1\}|$$

for all $g \in G$ (fixing $(s_1, ..., s_{k+1}) = (+1, ..., +1)$), so

$$(g(x_1), ..., g(x_k)) \neq (+1, ..., +1) = (s_1, ..., s_{k+1}) \in \{-1, +1\}^{k+1}$$

for all $g \in G$. This implies

$$|G_C| < 2^{k+1}$$

for all C of size $|C| = 2^{k+1}$, which completes the proof that all C of size $k+1$ are *not* shattered by G.
Since we already proved G shatters all sets C of size $|C| = k$, this completes the proof that the VC dimension of G is

$$VC(G) = k$$

3. Let $G = \{g_r, r \geq 0\}$ be the set of binary classifiers where $g_r : R^2 \to R$ is defined as

$$g_r(x) = \begin{cases} 1, & \|x\|_2 \leq r, \\ -1, & \|x\|_2 > r. \end{cases}$$

In other words, $g_r(x) = 1$ inside a circle of radius r. Show that $\text{VC}(G) = 1$.

Solution.
By the definition of VC dimension, to show

$$VC(G) := \max\{n \in \mathbb{N} : \tau_G(n) = 2^n\} = 1$$

we just need to:
i) Find a set $C = \{x_1\}$ of size $|C| = 1$ such that G shatters C (i.e. $|G_C| = 2^1 = 2$).
ii) Show that all sets $C = \{x_1, x_2\}$ of size $|C| = 2$ are *not* shattered by G (i.e. $|G_C| < 2^2 = 4$).
First, we will find a set $C = \{x_1\}$ of size $|C| = 1$ such that $|G_C| = 2$. Consider $C = \{x_1\} = \{(0, 2)\}$ and let $r_1 = 1$, $r_2 = 4$. Then $g_{r_1}(x_1) = -1$ because $||x_1||_2 = \sqrt{0^2 + 2^2} = 2 > r_1 = 1$ and $g_{r_2}(x_1) = 1$ because $||x_1||_2 = 2 \leq r_2 = 4$. So $(-1) \in G_C$ and $(+1) \in G_C$, so

$$G_C := \{(g(x_1), ..., g(x_n))|g \in G\} = \{(g(x_1))|g \in G\} = \{(-1), (+1)\}$$

so $|G_C| = 2$, which completes the proof that G shatters C.
Now, we will show all sets $C = \{x_1, x_2\}$ of size $|C| = 2$ are *not* shattered by G. Consider an arbitrary such $C = \{x_1, x_2\}$. It suffices to find a sequence $(s_1, s_2) \in \{-1, +1\}^2$ such that $(g(x_1), g(x_2)) \neq (s_1, s_2)$ for all $g \in G$.
If $||x_1||_2 = ||x_2||_2$, then $g_r(x_1) = g_r(x_2)$ for all $r \geq 0$ (i.e. for all $g \in G$), so $(g_r(x_1), g_r(x_2)) \in \{(+1, +1), (-1, -1)\}$. Thus, for all $g \in G$ (all $r \geq 0$), we have

$$(g_r(x_1), g_r(x_2)) \neq (-1, +1) = (s_1, s_2) \in \{-1, +1\}^2$$

If $||x_1||_2 < ||x_2||_2$, then $g_r(x_2) = +1 \implies r \geq ||x_2||_2 > ||x_1||_2 \implies g_r(x_1) = +1$. So $g_r(x_1) = -1, g_r(x_2) = +1$ is impossible. Thus, for all $g \in G$ (all $r \geq 0$), we have

$$(g_r(x_1), g_r(x_2)) \neq (-1, +1) = (s_1, s_2) \in \{-1, +1\}^2$$

Similarly, if $||x_2||_2 < ||x_1||_2$, then $g_r(x_1) = +1 \implies r \geq ||x_1||_2 > ||x_2||_2 \implies g_r(x_2) = +1$ for all $g \in G$ ($r \geq 0$). So $g_r(x_1) = +1, g_r(x_2) = -1$ is impossible. Thus, for all $g \in G$, we have

$$(g_r(x_1), g_r(x_2)) \neq (+1, -1) = (s_1, s_2) \in \{-1, +1\}^2$$

Regardless of the relative values of $||x_1||_2$ and $||x_2||_2$, we can always find a sequence $(s_1, s_2) \in \{-1, +1\}^2$ such that $(g(x_1), g(x_2)) \neq (s_1, s_2)$ for all $g \in G$. Thus, regardless of the composition of $C = \{x_1, x_2\}$, we always have

$$|G_C| < 4$$

which completes the proof that all sets C of size $|C| = 2$ are *not* shattered by G.
Since we already found a set C of size $|C| = 1$ that G *does* shatter, this completes the proof that $VC(G) = 1$.

4. Consider the set of all circles on the plane (not only the ones centered at the origin), and a set of binary classifiers that take value +1 inside the circles and -1 outside. We want to show that this class G of binary classifiers has VC dimension 3:

(a) First, consider the class G_1 of binary classifiers that take values +1 and -1 on half-planes corresponding to all lines of the form $\{(x, y) : \ ax + by + c = 0, \ a, b, c \in R, \text{ either } a \text{ or } b \neq 0\}$ (namely, half-planes are the regions above and below the lines). Show that VC dimension of G can not be larger than the VC dimension of G_1.

(hint: assume that a collection of points $M = \{x_1, \ldots, x_k\}$ is shattered by the circles. It means that for any subset M_1 of M, there exists a circle that contains only M_1 and another circle that contains only $M \setminus M_1$. Note that if these two circles intersect, there can not be any points from M in the intersection. Now try to find two half-planes with the same property - one that contains only M_1 and another - only $M \setminus M_1$. Drawing a picture will help!)

(b) Show, using the theorem we proved in class, that VC dimension of G_1 is 3 and deduce that $\text{VC}(G) \leq 3$.

(c) Find a set of 3 points shattered by G. State your conclusion.

(d) (*bonus) Now try to follow a similar approach to find the VC dimension of the class of binary classifiers that take value +1 inside the sphere in 3-dimensional space.

Solution.
First, define

$$L := \{f : \mathbb{R}^2 \to \mathbb{R} | f(x, y) = ax + by + c, \quad a, b, c \in \mathbb{R}\}$$

This definition will hold for parts (a), (b), and (c).

(a) We want to show that $VC(G) \leq VC(G_1)$. Note that

$$G := \{g_C | C \in R^2\}$$

where C is a circle in R^2 of the form

$$\{(x, y) \in R^2 | (x - a)^2 + (y - b)^2 \leq r^2 \quad a, b, r \in \mathbb{R}\}$$

and

$$g_C(x, y) := \begin{cases} +1 \text{ if } (x, y) \in C \\ -1 \text{ otherwise} \end{cases}$$

Consider any set of points $M \subseteq \mathbb{R}^2$, $M := \{m_1, ..., m_n\} = \{(x_1, y_1), ..., (x_n, y_n)\}$ such that G shatters M. This is true if and only if for all $M_1 \subseteq M$, $\exists$ a circle C_1 such that $g_{C_1}(m) = +1$ for all $m \in M_1$, $g_{C_1}(m) = -1$ for all $m \in M \setminus M_1$ and a circle C_2 such that $g_{C_2}(m) = +1$ for all $m \in M \setminus M_1$, $g_{C_2}(m) = -1$ for all $m \in M_1$. Since $g_C(m) = +1$ only when $m \in C$ by definition, we know G shatters $M \iff \exists C_1, C_2 \subseteq R^2$ s.t. $M_1 \subseteq C_1$, $(M \setminus M_1) \subseteq C_2$, and $C_1 \cap C_2 \cap M = \emptyset$ for all $M_1 \subseteq M$.
Claim: If $\exists C_1, C_2 \subseteq R^2$ s.t. $M_1 \subseteq C_1$, $(M \setminus M_1) \subseteq C_2$, and $C_1 \cap C_2 \cap M = \emptyset$ for all $M_1 \subseteq M$, then $\exists f(x, y) \in L$ such that $M_1 \subseteq \{(x, y) | f(x, y) > 0\}$, $M \setminus M_1 \subseteq \{(x, y) | -f(x, y) > 0\}$, and $M \cap \{(x, y) | f(x, y) > 0\} \cap \{(x, y) | -f(x, y) > 0\} = \emptyset$.
Proof. First, we note that

$$\begin{aligned} & M \cap \{(x, y) | f(x, y) > 0\} \cap \{(x, y) | -f(x, y) > 0\} \\ \subseteq \quad & \{(x, y) | f(x, y) > 0\} \cap \{(x, y) | -f(x, y) > 0\} \\ = \quad & \{(x, y) | f(x, y) > 0\} \cap \{(x, y) | f(x, y) < 0\} = \emptyset \end{aligned}$$

so $M \cap \{(x, y) | f(x, y) > 0\} \cap \{(x, y) | -f(x, y) > 0\} = \emptyset$ is trivially true for all f and all M since, for all $(x, y) \in \mathbb{R}^2$, $f(x, y) > 0 \implies f(x, y) \not< 0$. Thus, it suffices to show, If $\exists C_1, C_2 \subseteq R^2$ s.t. $M_1 \subseteq C_1$, $(M \setminus M_1) \subseteq C_2$, and $C_1 \cap C_2 \cap M = \emptyset$ for all $M_1 \subseteq M$, then $\exists f \in L$ such that $M_1 \subseteq \{(x, y) | f(x, y) > 0\}$, and $M \setminus M_1 \subseteq \{(x, y) | -f(x, y) > 0\}$. We will break this into three cases:

Case 1: $|C_1 \cap C_2| = 0$. Let (X_1, Y_1) be the center of C_1 and (X_2, Y_2) be the center of C_2.

Let r_1 = the radius of C_1 and r_2 = the radius of C_2. Consider the line through (X_1, Y_1) and (X_2, Y_2). If we write the line as

$$\{(x, y) \in \mathbb{R}^2 | y = mx + e\}$$

then we know $m = \frac{Y_2 - Y_1}{X_2 - X_1}$, so we can quickly compute that

$$Y_1 = \frac{Y_2 - Y_1}{X_2 - X_1} X_1 + e \implies e = Y_1 - \frac{Y_2 - Y_1}{X_2 - X_1} X_1$$

so the line that passes through (X_1, Y_1) and (X_2, Y_2) (the centers of both circles) is

$$L_1 := \{(x, y) \in \mathbb{R}^2 | y = \frac{Y_2 - Y_1}{X_2 - X_1} x + Y_1 - \frac{Y_2 - Y_1}{X_2 - X_1} X_1\}$$

Let $A = (X_1, Y_1)$ and $B = (X_2, Y_2)$. Consider the line segment $\overline{AB}$.

Claim: $|\overline{AB}| > r_1 + r_2$.
Proof: Assume to the contrary that $|AB| \leq r_1 + r_2$. Note that

$$|\overline{AB}| = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

. Fix a point $D = (X, Y)$ such that $|AC| = \sqrt{(X - X_1)^2 + (Y - Y_1)^2} = r_1$. Since $|\overline{AB}| = |\overline{AC}| + |\overline{CB}|$, we have

$$\begin{aligned} \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} = & \quad \sqrt{(X - X_1)^2 + (Y - Y_1)^2} + \sqrt{(X_2 - X)^2 + (Y_2 - Y)^2} \\ = & \quad r_1 + \sqrt{(X_2 - X)^2 + (Y_2 - Y)^2} \end{aligned}$$

which directly implies that

$$\sqrt{(X_2 - X)^2 + (Y_2 - Y)^2} = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} - r_1 \leq r_1 + r_2 - r_1 = r_2$$

with the last equality following by assumption.
For all $(x, y) \in \mathbb{R}^2$ such that $\sqrt{(x - X_1)^2 + (y - Y_1)^2} \leq r_1$, we know $(x, y) \in C_1$. Similarly, for all $(x, y) \in \mathbb{R}^2$ such that $\sqrt{(x - X_2)^2 + (y - Y_2)^2} \leq r_2$, we know $(x, y) \in C_2$. Thus, we have $(X, Y) \in C_1$ and $(X, Y) \in C_2$, so $(X, Y) \in C_1 \cap C_2$. But, by the definition of *Case 1*, we know $C_1 \cap C_2 = \emptyset \implies \forall (x, y) \in R^2 (x, y) \notin C_1 \cap C_2$, so we have a contradiction. This completes the proof that $|\overline{AB}| > r_1 + r_2$.

Since $|\overline{AB}| > r_1 + r_2$, we can split $\overline{AB}$ into three segments, $\overline{AD_1}$, $\overline{D_1 D_2}$, $\overline{D_2 B}$, such that $|\overline{AD_1}| = r_1$, $\overline{D_2 B} = r_2$, and $|\overline{D_1 D_2}| > 0$. Thus, for all points (x, y) on $\overline{D_2 D_1}$, we know (x, y) is more than r_1 from the center of C_1 and more than r_2 from the center of C_2, so $(x, y) \notin C_1 \cup C_2$. Arbitrarily pick such an (x^*, y^*) on $\overline{D_1 D_2}$ and consider the line orthogonal to $\overline{AB}$ that passes through (x^*, y^*). Call this line

$$L_2 := \{(x, y) \in \mathbb{R}^2 | y = m_2 x + e_2 \quad m_2, e_2 \in \mathbb{R}\}$$

Then $m_2 = -\frac{1}{m} = -\frac{X_2 - X_1}{Y_2 - Y_1}$ so we can easily compute

$$e_2 = y^* - m_2 x^* = y^* + \frac{X_2 - X_1}{Y_2 - Y_1} x^*$$

so

$$L_2 := \{(x, y) \in \mathbb{R}^2 | y = -x \frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1} x^*\}$$

Since L_2 and L_1 are orthogonal, the closest point on L_2 to C_1 is the closest point on L_2 to C_2 which is (x^*, y^*), the point of intersection of L_2 and L_1. Since $(x^*, y^*) \notin C_1 \cup C_2$, we know that for all $(x, y) \in L_2$, $(x, y) \notin C_1 \cup C_2$, so

$$\begin{aligned} C_1 \cup C_2 \subseteq & \quad \{(x,y) | y \neq -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\} \\ = & \quad \{(x,y) | y < -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\} \\ \cup & \quad \{(x,y) | y > -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\} \end{aligned}$$

Since L_2 intersects $\overline{AB}$, which passes between the center of C_1 and the center of C_2, we know

$$C_1 \subseteq \{(x,y) | y < -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

and

$$C_2 \subseteq \{(x,y) | y > -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

or

$$C_2 \subseteq \{(x,y) | y < -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

and

$$C_1 \subseteq \{(x,y) | y > -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

. Assume without loss of generality that

$$C_1 \subseteq \{(x,y) | y < -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

and

$$C_2 \subseteq \{(x,y) | y > -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

Then $M_1 \subseteq C_1 \implies$

$$M_1 \subseteq \{(x,y) | y < -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

and $M \setminus M_1 \subseteq C_2 \implies$

$$M \setminus M_1 \subseteq \{(x,y) | y > -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

which completes the proof for Case 1.

Case 2: $|C_1 \cap C_2| = 1$. Define L_1, (X_1, Y_1), (X_2, Y_2), $\overline{AB}$, r_1 and r_2 as in *Case 1.* Then redefine (x^*, y^*) such that $\{(x^*, y^*)\} = C_1 \cap C_2$, and the line orthogonal to L_1 that passes through (x^*, y^*) is L_2 as defined before (now in terms of the redefined (x^*, y^*)). Once again, since L_2 is orthogonal to L_1, we know the closest point on L_2 to C_1 is the closest point on L_2 to C_2 which is (x^*, y^*) (the intersection of L_1 and L_2). Since $\{(x^*, y^*)\} = C_1 \cap C_2$, we know (x^*, y^*) is on the boundary of both C_1 and C_2 (otherwise $|C_1 \cap C_2| > 1 \implies \{(x^*, y^*)\} \neq C_1 \cap C_2$). Thus, $\sqrt{(X_1 - x^*)^2 + (Y_1 - y^*)^2} = r_1$ and $\sqrt{(X_2 - x^*)^2 - (Y_2 - y^*)^2} = r_2$, so for all $(x, y) \in L_2$ such that $(x, y) \neq (x^*, y^*)$, we have $\sqrt{(X_1 - x)^2 + (Y_1 - y)^2} > r_1$ and $\sqrt{(X_2 - x)^2 + (Y_2 - y)^2} > r_2$, so $(x, y) \notin C_1 \cup C_2$. Since L_2 only intersects C_1 at $\{(x^*, y^*)\} = C_1 \cap C_2$ and L_2 only intersects C_2 at $\{(x^*, y^*)\} = C_1 \cap C_2$, and L_2 intersects $\overline{AB}$ we know either

$$C_1 \cap C_2^c \subseteq \{(x,y) | y < -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

and

$$C_2 \cap C_1^c \subseteq \{(x,y)|y > -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

or

$$C_2 \cap C_1^c \subseteq \{(x,y)|y < -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

and

$$C_1 \cap C_2^c \subseteq \{(x,y)|y > -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

. Assume without loss of generality that

$$C_1 \cap C_2^c \subseteq \{(x,y)|y < -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

and

$$C_2 \cap C_1^c \subseteq \{(x,y)|y > -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

Then $M_1 \subseteq C_1$, $M \cap C_1 \cap C_2 = \emptyset$, $M_1 \subseteq M \implies M_1 \subseteq C_1 \cap C_2^c$, which implies

$$M_1 \subseteq \{(x,y)|y < -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

and $M \setminus M_1 \subseteq C_2$, $M \cap C_1 \cap C_2 = \emptyset$, $M \setminus M_1 \subseteq M \implies M \setminus M_1 \subseteq C_2 \cap C_1^c$

$$M \setminus M_1 \subseteq \{(x,y)|y > -x\frac{X_2 - X_1}{Y_2 - Y_1} + y^* + \frac{X_2 - X_1}{Y_2 - Y_1}x^*\}$$

which completes the proof for Case 2.

Case 3: $|C_1 \cap C_2| > 1$. Define (X_1, Y_1), (X_2, Y_2), L_1, $\overline{AB}$, r_1, and r_2 as before. When two circles intersect at more than one point, their borders intersect at exactly 2 points. Call these two points (x_a, y_a) and (x_b, y_b). Then we can redefine L_2 to be the line that passes through these two points, and L_2 will still be orthogonal to L_1. We can apply the same computations we used in *Case 1* to find L_1, this time with (x_a, y_a) and (x_b, y_b) instead of (X_1, Y_1) and (X_2, Y_2) to find

$$L_2 = \{(x,y)|y = \frac{y_b - y_a}{x_b - x_a}x + y_a - \frac{y_b - y_a}{x_b - x_a}x_a\}$$

Since L_2 passes through both points in the intersection of the borders of C_1 and C_2, all points in C_1 but not C_2 must be on one side of L_2. That is, either

$$C_1 \cap C_2^c \subseteq \{(x,y)|y < \frac{y_b - y_a}{x_b - x_a}x + y_a - \frac{y_b - y_a}{x_b - x_a}x_a\}$$

and

$$C_2 \cap C_1^c \subseteq \{(x,y)|y > \frac{y_b - y_a}{x_b - x_a}x + y_a - \frac{y_b - y_a}{x_b - x_a}x_a\}$$

or

$$C_2 \cap C_1^c \subseteq \{(x,y)|y < \frac{y_b - y_a}{x_b - x_a}x + y_a - \frac{y_b - y_a}{x_b - x_a}x_a\}$$

and

$$C_1 \cap C_2^c \subseteq \{(x,y)|y > \frac{y_b - y_a}{x_b - x_a}x + y_a - \frac{y_b - y_a}{x_b - x_a}x_a\}$$

. Assume without loss of generality that

$$C_1 \cap C_2^c \subseteq \{(x,y)|y < \frac{y_b - y_a}{x_b - x_a}x + y_a - \frac{y_b - y_a}{x_b - x_a}x_a\}$$

and

$$C_2 \cap C_1^c \subseteq \{(x,y)|y > \frac{y_b - y_a}{x_b - x_a}x + y_a - \frac{y_b - y_a}{x_b - x_a}x_a\}$$

Then by the exact same logic as in *Case 2*, we have

$$M_1 \subseteq \{(x,y)|y < \frac{y_b - y_a}{x_b - x_a}x + y_a - \frac{y_b - y_a}{x_b - x_a}x_a\}$$

and

$$M \setminus M_1 \subseteq \{(x,y)|y > \frac{y_b - y_a}{x_b - x_a}x + y_a - \frac{y_b - y_a}{x_b - x_a}x_a\}$$

which completes the proof for *Case 3*.

Thus, regardless of the size of the intersection of C_1 and C_2, we can always find two half planes A and B of the form $\{(x,y)|y < mx + e\}$ and $\{(x,y)|y > mx + e\}$ respectively such that

$$M_1 \subseteq A, \quad M \setminus M_1 \subseteq B, \quad M \cap A \cap B = \emptyset$$

for all $M_1 \subseteq M$.
If we choose $f_1, f_2 \in L$ s.t.

$$f_1(x,y) := mx + e - y = mx - y + e \quad f_2(x,y) := y - mx - e = -mx + y - e$$

then $f_1(x,y) > 0 \iff y < mx + e$ and $f_2(x,y) > 0 \iff y > mx + e$, so

$$\{(x,y)|f_1(x,y) > 0\} = \{(x,y)|y < mx + e\} = A$$

and

$$\{(x,y)|f_2(x,y) > 0\} = \{(x,y)|y > mx + e\} = B$$

By definition, the corresponding classifiers $g_{f_1}, g_{f_2} \in G_1$ satisfy

$$g_{f_1}(m) = +1, \text{ for all } m \in M_1 \quad g_{f_2}(m) = +1 \text{ for all } m \in M \setminus M_1$$

for all $M_1 \subseteq M$. Note that

$$\begin{array}{cl} & M \cap \{(x,y)|f_1(x,y) > 0\} \cap \{(x,y)|f_2(x,y) > 0\} \\ \subseteq & \{(x,y)|f_1(x,y) > 0\} \cap \{(x,y)|f_2(x,y) > 0\} \\ = & \{(x,y)|f_1(x,y) > 0\} \cap \{(x,y)| - f_1(x,y) > 0\} \\ = & \{(x,y)|f_1(x,y) > 0\} \cap \{(x,y)|f_1(x,y) < 0\} = \emptyset \end{array}$$

implies

$$M \cap \{(x,y)|f_1(x,y) > 0\} \cap \{(x,y)|f_2(x,y) > 0\} = \emptyset$$

which implies

$$g_{f_1}(m) = -1, \text{ for all } m \in M \setminus M_1 \quad g_{f_2}(m) = +1 \text{ for all } m \in M_1$$

for all $M_1 \subseteq M$. Thus, for any of the 2^n subsets $M_1 \subseteq M$, we can find a $g_f \in G_1$ s.t. $g_f(m) = +1$ for all $m \in M_1$ and $g_f(m) = -1$ for all $m \in M \setminus M_1$. This holds for all M under our initial assumption that G shatters M. Thus, for any collection of points M of size n such that G shatters M, we can produce all 2^n sequences of signs $(s_1, ..., s_n) \in \{-1, +1\}^n$ with sequences of the form $(g(m_1), ..., g(m_n))$, where $g \in G_1$. Thus, $\tau_{G_1}(n) = 2^n$ for any such M, so any M shattered by G is also shattered by G_1.
From here, it is simple to show $VC(G) \leq VC(G_1)$. Assume to the contrary that $VC(G) > VC(G_1)$. Then, by the definition of VC dimension, we have

$$\max\{n \in \mathbb{N}|\tau_G(n) = 2^n\} > \max\{n \in \mathbb{N}|\tau_{G_1}(n) = 2^n\}$$

so if we let $M :=$ the biggest collection of points such that G shatters M and $M_1 :=$ the biggest collection of points such that G_1 shatters M_1, then we have

$$|M| > |M_1|$$

But G shatters M implies G_1 shatters M, so G_1 shatters a set M larger than M_1. By the definition of M_1, we have a contradiction. Thus, our initial assumption that $VC(G) > VC(G_1)$ must be incorrect, which completes the proof that

$$VC(G) \leq VC(G_1)$$

(b) We want to show that $VC(G_1) = 3$. First, recall

$$L := \{f : \mathbb{R}^2 \to \mathbb{R} | f(x, y) = ax + by + c, \quad a, b, c \in \mathbb{R}\}$$

and consider the set

$$G^* := \{g_f | f \in L\}$$

where

$$g_f(x, y) := \begin{cases} +1 \text{ if } f(x, y) > 0 \\ -1 \text{ otherwise} \end{cases}$$

First, note that, for all $f \in L$ such that $f(x, y) = 0x + 0y + c = c$, $g_f \notin G_1$. However, for all such f where $c > 0$, we have g_f identically equal to 1, as $f(x, y) = c > 0$ over all of $\mathbb{R}^2$. Similarly, for all such f where $c \leq 0$, we have g_f identically equal to -1, as $f(x, y) = c \leq 0$ over all of $\mathbb{R}^2$. From the problem statement's definition of G_1, we know that for $f(x, y) = \lim_{a,b \to 0, c \to \infty} ax+by+c$, $g_f \in G_1$. By setting a and b to be arbitrarily small while c is arbitrarily large and positive, we guarantee that $f(x, y) \approx c > 0$ for all $(x, y) \in \mathbb{R}^2$. Similarly, we know that for $f(x, y) = \lim_{a,b \to 0, c \to -\infty} ax+by+c$, $g_f \in G_1$. Here, since c is arbitrarily large and negative while a and b approach 0, we have $f(x, y) \approx c \leq 0$, so g_f is identically -1 over all of $\mathbb{R}^2$. Thus, although $\exists f \in L$ such that $g_f \notin G_1$, for all collections of points $M = \{x_1, ..., x_n\} \subseteq \mathbb{R}^2$, no such g_f produces a sequence $(g_f(x_1), ..., g_f(x_n))$ which cannot be produced g where $g \in G_1$. Therefore, all elements of

$$G^*_M := \{(g(x_1), ..., g(x_n)) | g \in G^*\}$$

produced by $g_f \notin G_1$ can also be produced by different $g_f \in G_1$. This implies that

$$|G^*_M| \leq |(G_1)_M|$$

for all collections M, so $\tau_{G^*}(n) \leq \tau_{G_1}(n)$ for all $n \in \mathbb{N}$. By the definition of VC dimension, we have

$$VC(G_1) = \max\{n \in \mathbb{N} : \tau_{G_1}(n) = 2^n\} \geq \max\{n \in \mathbb{N} : \tau_{G^*}(n) = 2^n\} := VC(G^*)$$

But, as it is defined in the problem statement,

$$\begin{aligned} G_1 = \quad & \{g_f | f \in \{f : \mathbb{R}^2 \to \mathbb{R} | f(x, y) = ax + by + c, \text{ either } a \text{ or } b \neq 0\}\} \\ \subseteq \quad & \{g_f | f \in L\} := G^* \end{aligned}$$

because

$$\begin{aligned} & \{f : \mathbb{R}^2 \to \mathbb{R} | f(x, y) = ax + by + c, \text{ either } a \text{ or } b \neq 0\} \\ \subseteq \quad & \{f : \mathbb{R}^2 \to \mathbb{R} | f(x, y) = ax + by + c, \quad a, b, c \in \mathbb{R}\} := L \end{aligned}$$

Since $G_1 \subseteq G^*$, we know that for all collections $M = \{x_1, ..., x_n\}$, if $g \in G_1$, $g \in G^*$, so $(g(x_1), ..., g(x_n)) \in G^*_M$ for all $(g(x_1), ..., g(x_n)) \in (G_1)_M$. This implies that

$$(G_1)_M \subseteq G^*_M \implies |(G_1)_M| \leq |G^*_M|$$

for all collections $M = \{x_1, ..., x_n\}$ which implies

$$\tau_{G_1}(n) \leq \tau_{G^*}(n)$$

for all $n \in \mathbb{N}$. By the definition of VC dimension, we have

$$VC(G_1) := \max\{n \in \mathbb{N} : \tau_{G_1}(n) = 2^n\} \leq \max\{n \in \mathbb{N} : \tau_{G^*}(n) = 2^n\} := VC(G^*)$$

Since $VC(G_1) \leq VC(G^*)$ and $VC(G_1) \geq VC(G^*)$, we know

$$VC(G_1) = VC(G^*)$$

Now, to show $VC(G_1) = 3$, it suffices to show $VC(G^*) = 3$.
Consider two arbitrary functions $f_1, f_2 \in L$ and pick $a_1, a_2, b_1, b_2, c_1, c_2 \in \mathbb{R}$ such that

$$f_1(x, y) = a_1x + b_1y + c_1 \quad f_2(x, y) = a_2x + b_2y + c_2$$

Note that $f_1, f_2 \in L$. Consider two arbitrary constants $u, v \in R$. Then

$$\begin{aligned} uf_1(x, y) + vf_2(x, y) = & \quad u(a_1x + b_1y + c_1) + v(a_2x + b_2y + c_2) \\ = & \quad ua_1x + ub_1y + uc_1 + va_2x + vb_2y + vc_2 \\ = & \quad (ua_1 + va_2)x + (ub_1 + vb_2)y + (uc_1 + vc_2) \end{aligned}$$

Since $u, v, a_1, a_2, b_1, b_2, c_1, c_2 \in \mathbb{R}$, we have $(ua_1 + va_2), (ub_1 + vb_2), (uc_1 + vc_2) \in R$. Thus, we can let $a = ua_1 + va_2$, $b = ub_1 + vb_2$, and $c = uc_1 + vc_2$, and we find

$$uf_1(x, y) + vf_2(x, y) = ax + by + c = f_3(x, y) \in L$$

Thus, L is closed under both addition and scalar multiplication, so L is a linear space.
The R. Dudley Theorem from lecture tells us that for any linear space L of functions $f : S \to \mathbb{R}$ with finite dimension $dim(L) = d$, the family of classifiers $G := \{g_f | f \in L\}$ where

$$g_f(x) := \begin{cases} +1 \text{ if } f(x) > 0 \\ -1 \text{ if } f(x) \leq 0 \end{cases}$$

for all $x \in S$ has VC dimension

$$VC(G) = dim(L) = d$$

Here $S = \mathbb{R}^2$, $G^* := \{g_f | f \in L\}$, L is a linear space, and

$$g_f(x, y) := \begin{cases} +1 \text{ if } f(x, y) > 0 \\ -1 \text{ if } f(x, y) \leq 0 \end{cases}$$

so we can apply the R. Dudley Theorem to find $VC(G^*)$. All we have to do is prove that L has finite dimension and compute $dim(L)$. From here, we can conclude

$$VC(G^*) = dim(L)$$

By definition, the dimension of a linear space L is the size of any of its bases. Since

$$L := \{f : \mathbb{R}^2 \to \mathbb{R} | f(x, y) = ax + by + c, a, b, c \in \mathbb{R}\}$$

we can construct every function in L with linear combinations of the vectors (x), (y), and (1) (i.e. $\{x, y, 1\}$ spans L). Furthermore, none of these vectors can be written as a linear combination of the other two

$$\begin{aligned} &x \neq uy + v(1) \text{ for all } u, v \in \mathbb{R} \\ &y \neq ux + v(1) \text{ for all } u, v \in \mathbb{R} \\ &1 \neq ux + vy \text{ for all } u, v \in \mathbb{R} \end{aligned}$$

By definition of the basis of a linear space, as a linearly independent spanning set of L, $\{x, y, 1\}$ is a basis for L. Thus, by the definition of the dimension of a linear space, we have

$$dim(L) = |\{x, y, 1\}| = 3$$

Thus, by the R. Dudley Theorem from lecture, we have

$$VC(G^*) = dim(L) = 3$$

We already proved $VC(G^*) = VC(G_1)$, so this completes the proof that

$$VC(G_1) = 3$$

Since we proved in part (a) that $VC(G) \leq VC(G_1)$, we can conclude

$$VC(G) \leq 3$$

which completes part (b).

(c) *Claim:* $M = \{x_1, x_2, x_3\}\{(0,0), (2,0), (2,2)\}$ is a set of 3 points shattered by G.
Proof. It suffices to show that $|G_M| = 8$, so we need to find 8 distinct $g \in G$ to produce all 8 distinct $(s_1, s_2, s_3) \in \{-1, +1\}^3$. For any circle $C := \{(x, y)|(x-a)^2 + (y-b)^2 \leq r^2, a, b, r \in \mathbb{R}\}$, we can write the corresponding classifier $g_C \in G$ as

$$g_C(x, y) := \begin{cases} +1 \text{ if } (x, y) \in C \\ -1 \text{ if } (x, y) \notin C \end{cases}$$

Then let

$$\begin{aligned}
C_1 &:= \{(x, y)|x^2 + y^2 \leq 16\} \\
C_2 &:= \{(x, y)|(x+2)^2 + y^2 \leq 1\} \\
C_3 &:= \{(x, y)|x^2 + y^2 \leq 1\} \\
C_4 &:= \{(x, y)|(x-2)^2 + y^2 \leq 1\} \\
C_5 &:= \{(x, y)|(x-2)^2 + (y-2)^2 \leq 1\} \\
C_6 &:= \{(x, y)|(x-1)^2 + (y+1)^2 \leq 5\} \\
C_7 &:= \{(x, y)|(x-3)^2 + (y-1)^2 \leq 5\} \\
C_8 &:= \{(x, y)|x^2 + (y-2)^2 \leq 5\}
\end{aligned}$$

and consider the corresponding classifiers $g_{C_1}, ..., g_{C_8} \in G$. We find

$$(g_{C_1}(x_1), g_{C_1}(x_2), g_{C_1}(x_3)) = (+1, +1, +1)$$

since$0^2 + 0^2 = 0 \leq 16, 2^2 + 0^2 = 4 \leq 16, 2^2 + 2^2 = 8 \leq 16$,

$$(g_{C_2}(x_1), g_{C_2}(x_2), g_{C_2}(x_3)) = (-1, -1, -1)$$

since$(0+2)^2 + 0^2 = 4 \not\leq 1, (2+2)^2 + 0^2 = 16 \not\leq 1, (2+2)^2 + 2^2 = 20 \not\leq 1$,

$$(g_{C_3}(x_1), g_{C_3}(x_2), g_{C_3}(x_3)) = (+1, -1, -1)$$

since $0^2 + 0^2 = 0 \leq 1, 2^2 + 0^2 = 4 \not\leq 1, 2^2 + 2^2 = 8 \not\leq 1$,

$$(g_{C_4}(x_1), g_{C_4}(x_2), g_{C_4}(x_3)) = (-1, +1, -1)$$

since $(0-2)^2 + 0^2 = 4 \not\leq 1, (2-2)^2 + 0^2 = 0 \leq 1, (2-2)^2 + 2^2 = 4 \not\leq 1$,

$$(g_{C_5}(x_1), g_{C_5}(x_2), g_{C_5}(x_3)) = (-1, -1, +1)$$

since $(0-2)^2+(0-2)^2=8 \not\leq 1, (2-2)^2+(0-2)^2=4 \not\leq 1, (2-2)^2+(2-2)^2=0 \leq 1$,

$$(g_{C_6}(x_1), g_{C_6}(x_2), g_{C_6}(x_3)) = (+1, +1, -1)$$

since $(0-1)^2+(0+1)^2=2 \leq 5, (2-1)^2+(0+1)^2=2 \leq 5, (2-1)^2+(2+1)^2=10 \not\leq 5$,

$$(g_{C_7}(x_1), g_{C_7}(x_2), g_{C_7}(x_3)) = (-1, +1, +1)$$

since $(0-3)^2+(0-1)^2=10 \not\leq 5, (2-3)^2+(0-1)^2=2 \leq 5, (2-3)^2+(2-1)^2=2 \leq 5$,

$$(g_{C_8}(x_1), g_{C_8}(x_2), g_{C_8}(x_3)) = (+1, -1, +1)$$

since $0^2+(0-2)^2=4 \leq 5, 2^2+(0-2)^2=8 \not\leq 5, 2^2+(2-2)^2=4 \leq 5$. Thus, $g_{C_1}, ..., g_{C_8}$ are 8 distinct $g \in G$ that combine to produce all 8 distinct $(s_1, s_2, s_3) \in \{-1, +1\}^3$, so we have

$$|G_M| := |\{(g(x_1), g(x_2), g(x_3)) | g \in G\}| = 8$$

so $\tau_G(3) = 2^3 = 8$, so G shatters M. By definition, the VC dimension of G is

$$VC(G) := \max\{n \in \mathbb{N} : \tau_G(n) = 2^n\}$$

Since $\tau_G(3) = 2^3$, we know $3 \in \{n \in \mathbb{N} : \tau_G(n) = 2^n$, so we know

$$VC(G) \geq 3$$

However, we already showed in parts (a) and (b) that $VC(G) \leq VC(G_1) = VC(G^*) = 3$, so we can conclude

$$VC(G) = 3$$

This completes the proof that the VC dimension of G is 3.

(d) (*bonus) Now, let G be the class of binary classifiers that take value $+1$ inside the sphere in 3-dimensional space. We want to find $VC(G)$.
Claim: The VC dimension of G is $VC(G) = 4$.
Proof. We will follow a very similar proof to that from parts (a), (b), and (c).
First, define

$$L := \{f : \mathbb{R}^3 \to \mathbb{R} | f(x, y, z) = ax + by + cz + d\}$$

where $a, b, c, d \in \mathbb{R}$ are scalars. Arbitrarily pick two functions $f_1, f_2 \in L$ and eight scalars $a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2 \in \mathbb{R}$ such that

$$f_1(x, y, z) := a_1x + b_1y + c_1z + d_1 \quad f_2(x, y, z) := a_2x + b_2y + c_2z + d_2$$

Then choose two more arbitrary scalars $u, v \in \mathbb{R}$, and we find

$$\begin{aligned} uf_1(x, y, z) + vf_2(x, y, z) := &\quad u(a_1x + b_1y + c_1z + d_1) + v(a_2x + b_2y + c_2z + d_2) \\ = &\quad ua_1x + ub_1y + uc_1z + ud_1 + va_2x + vb_2y + vc_2z + vd_2 \\ = &\quad (ua_1 + va_2)x + (ub_1 + vb_2)y + (uc_1 + vc_2)z + (ud_1 + vd_2) \end{aligned}$$

Since $a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2, u, v \in \mathbb{R}$, we know $(ua_1+va_2), (ub_1+vb_2), (uc_1+vc_2), (ud_1+vd_2) \in R$, so we can let $a = (ua_1 + va_2), b = (ub_1 + vb_2), c = (uc_1 + vc_2), d = (ud_1 + vd_2)$ and we have

$$uf_1(x, y, z) + vf_2(x, y, z) = ax + by + cz + d = f_3(x, y, z) \in L$$

Thus, L is closed under addition and scalar multiplication, so we know L is a linear space. Define

$$G_1 := \{g_f | f \in L\}$$

where

$$g_f(x,y,z) := \begin{cases} +1 \text{ if } f(x,y,z) > 0 \\ -1 \text{ if } f(x,y,z) \leq 0 \end{cases}$$

By the R. Dudley Theorem from lecture, if the dimension of L is finite (i.e. $dim(L) < \infty$), then we have

$$VC(G_1) = dim(L)$$

By definition of the dimension of a linear space, we know the dimension of L is equal to the size of any of its bases. Note that for all $ax+by+cz+d = f(x,y,z) \in L$, $f(x,y,z)$ is a linear combination of the vectors (1), (x), (y), and (z), so $\{1,x,y,z\}$ spans L. Also, none of these vectors can be produced by linear combinations of the others, i.e.

$$\begin{aligned} 1 &\neq ax + by + cd && \forall a,b,c \in \mathbb{R} \\ x &\neq ay + bd + c(1) && \forall a,b,c \in \mathbb{R} \\ y &\neq ax + bd + c(1) && \forall a,b,c \in \mathbb{R} \\ z &\neq ax + by + c(1) && \forall a,b,c \in \mathbb{R} \end{aligned}$$

Thus, $\{1,x,y,z\}$ is a linearly independent spanning set for L, so by the definition of a basis of a linear space, $\{1,x,y,z\}$ is a basis for L. Thus, by the definition of the dimension of a linear space, we have

$$dim(L) = |\{1,x,y,z\}| = 4$$

Thus, by the R. Dudley Theorem from lecture, since $dim(L) = 4 < \infty$, we have

$$VC(G_1) = dim(L) = 4$$

Next, we want to show that $VC(G) \leq VC(G_1)$.
Note that any 3-dimensional plane takes the form

$$P = \{(x,y,z) | ax + by + cz + d = 0, \quad \text{either } a,b \text{ or } c \neq 0\}$$

so any 3 dimensional half-plane (the region above or below a 3 dimensional plane) takes the form

$$H_{below} = \{(x,y,z) | ax + by + cz + d < 0, \quad \text{either } a,b \text{ or } c \neq 0\}$$

or

$$H_{above} = \{(x,y,z) | ax + by + cz + d > 0, \quad \text{either } a,b \text{ or } c \neq 0\}$$

Since

$$\{f : \mathbb{R}^3 \to \mathbb{R} | f(x,y,z) = ax + by + cz + d \quad \text{either } a,b \text{ or } c \neq 0\} \subseteq L$$

for all classifiers g whose sets of positivity include only a 3-dimensional half plane, which takes the form

$$\{(x,y,z) \in \mathbb{R}^3 | f(x,y,z) > 0\} \text{ or } \{(x,y,z) \in \mathbb{R}^3 | -f(x,y,z) > 0\}$$

we have $g \in G_1$.
Also, for any set $M := \{x_1, ..., x_n\} \subseteq \mathbb{R}^3$ of points in 3 dimensional space, G shatters $M \iff$ we have $g \in G$ such that

$$g(x) = \begin{cases} +1 \text{ for all } x \in M_1 \\ -1 \text{ for all } x \in M \setminus M_1 \end{cases}$$

for any arbitrary subset $M_1 \subseteq M$ (so M_1 could be the empty set or M itself). Since all $g \in G$ take value $+1$ only inside a sphere in 3-dimensional space, we know that we must have two spheres

$$\begin{aligned} S_1 &:= \{(x,y,z) | (x-a_1)^2 + (y-b_1)^2 + (z-c_1)^2 \leq r_1^2\} \\ S_2 &:= \{(x,y,z) | (x-a_2)^2 + (y-b_2)^2 + (z-c_2)^2 \leq r_2^2\} \end{aligned}$$

such that $M_1 \subseteq S_1$, $M \setminus M_1 \subseteq S_2$, and $M \cap S_1 \cap S_2 = \emptyset$. We showed that all classifiers g that take value $+1$ only on a 3 dimensional half-plane are all of the form $g_f \in G_1$. Thus, to prove that G_1 shatters any collection of points M shattered by G, we just need to show that, for any collection of points M and any of its subsets $M_1 \subseteq M$, if $\exists S_1, S_2 \subseteq \mathbb{R}^3$ such that $M_1 \subseteq S_1$, $M \setminus M_1 \subseteq S_2$, and $M \cap S_1 \cap S_2 = \emptyset$, then $\exists$ 3 dimensional half-planes H_1, H_2 such that $M_1 \subseteq H_1$, $M \setminus M_1 \subseteq H2$, and $M \cap H_1 \cap H_2 = \emptyset$. We will do so by considering the different sizes of $|S_1 \cap S_2|$, just like in part (a).

Case 1: $|S_1 \cap S_2| = 0$. Then by the same logic as in part (a), we know there is a point $C = (x, y, z) \in \mathbb{R}^3$ such that $(x, y, z) \notin S_1 \cup S_2$, but $(x, y, z) \in \overline{AB}$, where $A = (a_1, b_1, c_1)$ is the center of S_1 and $B = (a_2, b_2, c_2)$ is the center of S_2. Now, consider the 3 dimensional plane P drawn through C such that P is orthogonal to $\overline{AB}$. Since P is orthogonal to $\overline{AB}$, we know the closest point on P to $\overline{AB}$ is its point of intersection with $\overline{AB}$, which is defined to be C. Since $C \notin S_1 \cup S_2$, we know $|\overline{AC}| > r_1$ and $|\overline{CB}| > r_2$, so we know $P \cap (S_1 \cup S_2) = \emptyset$. Since P passes through $\overline{AB}$, which connects the centers of S_1 and S_2, but P doesn't pass through S_1 or S_2, we know all of S_1 is above P and all of S_2 is below P or all of S_1 is below P and all of S_2 is above P. Assume without loss of generality that all of S_1 is above P and all of S_2 is below P. If we let H_1 be the half plane above P and H_2 be the half plane below P, then

$$S_1 \subseteq H_1 \qquad S_2 \subseteq H_2$$

Since $M_1 \subseteq S_1 \subseteq H_1$ and $M \setminus M_1 \subseteq S_2 \subseteq H_2$, we have found two 3 dimensional half planes H_1, H_2 such that $M_1 \subseteq H_1$ and $M \setminus M_1 \subseteq H_2$. Also, since H_1 and H_2 are defined to be regions on opposite sides of P, we know that $H_1 \cap H_2 = \emptyset$, which combines with the fact that $M \cap H_1 \cap H_2 \subseteq H_1 \cap H_2$ to imply that $M \cap H_1 \cap H_2 = \emptyset$. This completes the proof for *Case 1.*

Case 2: $|S_1 \cap S_2| = 1$. Just like in part (a), let $C = (x, y, z)$ be the point of intersection between S_1 and S_2, and define A, B as before. Once again, consider the plane P that passes through the point C and is orthogonal to $\overline{AB}$. Since P is orthogonal to $\overline{AB}$, which connects the centers of S_1 and S_2, we know the closest point on P to S_1 and S_2 is P's intersection with $\overline{AB}$, which is C. Since $\{C\} = S_1 \cap S_2$ by definition, we know $C \in S_1$ and $C \in S_2$. Also, we know C is on the boundary of both S_1 and S_2 ($S_1 \cap S_2 \neq \emptyset$ implies the existence of at least one point d on the boundary of both S_1 and S_2. If this point $d \neq C$, then $d, C \in S_1 \cap S_2 \implies |S_1 \cap S_2| > 1$, which is a contradiction by the definition of the case). Thus, $|\overline{AC}| = r_1$ and $|\overline{BC}| = r_2$. Thus, since all points $e \in P$ such that $e \neq C$ are further from both S_1 and S_2 than C, we know $P \cap (S_1 \cup S_2) = \{C\}$. Since P intersects $\overline{AB}$, which connects the centers of S_1 and S_2, and P doesn't intersect S_1 or S_2 anywhere but C, where $\{C\} = S_1 \cap S_2$, we know all of $S_1 \cap S_2^c$ is on one side of P and all of $S_2 \cap S_1^c$ is on the other side of P. That is, if we let H_1 be the 3 dimensional half-plane above P and H_2 be the 3-dimensional half-plane below P, then either $S_1 \cap S_2^c \subseteq H_1$ and $S_2 \cap S_1^c \subseteq H_2$, or $S_1 \cap S_2^c \subseteq H_2$ and $S_2 \cap S_1^c \subseteq H_1$. Assume without loss of generality that $S_1 \cap S_2^c \subseteq H_1$ and $S_2 \cap S_1^c \subseteq H_2$. Then $M_1 \subseteq S_1 \subseteq H_1$ and $M \setminus M_1 \subseteq S_2 \subseteq H_2$ implies

$$M_1 \subseteq H_1 \qquad M \setminus M_1 \subseteq H_2$$

Also, since H_1 and H_2 are defined to be 3 dimensional half planes on *opposite* sides of the same plane P, we know $H_1 \cap H_2 = \emptyset$, so we know $M \cap H_1 \cap H_2 = \emptyset$ since $M \cap H_1 \cap H_2 \subseteq H_1 \cap H_2$. So we have found two 3 dimensional half planes H_1 and H_2 such that $M_1 \subseteq H_1$, $M \setminus M_1 \subseteq H_2$, and $M \cap H_1 \cap H_2 = \emptyset$. This completes the proof for *Case 2.*

Case 3: $|S_1 \cap S_2| > 1$. Note that whenever two spheres intersect at more than one point, the intersection of their boundaries forms the boundary of a circle. Define C to be the circle whose boundary is formed by the intersections of the boundaries of S_1 and S_2. Note that, as a 2 dimensional surface, C is contained entirely within a plane P, which is still orthogonal to $\overline{AB}$, defining A, B as before. Also, all of the points on C's boundary are on the boundaries of S_1 and

S_2, so all such points C_i are such that $|\overline{C_iA}| = r_1$ and $|\overline{C_iB} = r_2|$ All points on P that are closer to $\overline{AB}$ than any C_i on the boundary of C are also closer to both S_1 and S_2 since $\overline{AB}$ connects the centers of S_1 and S_2. Thus, all points on the interior of C are in $S_1 \cap S_2$. Conversely, all points on P that are further from $\overline{AB}$ than any C_i on the boundary of C are also further from both S_1 and S_2 since P is orthogonal to $\overline{AB}$, so we know all points in P that are in S_1 or S_2 are also in both S_1 and S_2. That is,

$$P \cap (S_1 \cup S_2) = P \cap S_1 \cap S_2 \quad P \cap (S_1 \cap S_2^c) = \emptyset \quad P \cap (S_2 \cap S_1^c) = \emptyset$$

Since P orthogonally intersects the line segments $\overline{AB}$ that connects the centers of S_1 and S_2, we know all of $S_1 \cap S_2^c$ is contained entirely one side of P and $S_2 \cap S_1^c$ is contained entirely on the other side of P. That is, if we let $H_1 :=$ the 3 dimensional half plane above P and $H_2 :=$ the 3 dimensional half plane below P, then either $S_1 \cap S_2^c \subseteq H_1$ and $S_2 \cap S_1^c \subseteq H_2$ or $S_1 \cap S_2^c \subseteq H_2$ and $S_2 \cap S_1^c \subseteq H_1$. Assume, without loss of generality, that $S_1 \cap S_2^c \subseteq H_1$ and $S_2 \cap S_1^c \subseteq H_2$. Then $M_1 \subseteq S_1 \subseteq H_1$ and $M \setminus M_1 \subseteq S_2 \subseteq H_2$ implies

$$M_1 \subseteq H_1 \text{ and } M \setminus M_1 \subseteq H_2$$

Also, since H_1 and H_2 are defined to be 3 dimensional half planes on *opposite* sides of the same plane P, we know $H_1 \cap H_2 = \emptyset$, so $M \cap H_1 \cap H_2 \subseteq H_1 \cap H_2$ implies that

$$M \cap H_1 \cap H_2 = \emptyset$$

Thus, we have found two 3 dimensional half-planes H_1 and H_2 such that $M_1 \subseteq H_1$, $M \setminus M_1 \subseteq H_2$, and $M \cap H_1 \cap H_2 = \emptyset$, which completes the proof for *Case 3.*

Thus, for any collection of points $M = \{x_1, ..., x_k\}$ that is shattered by G, we can find H_1, H_2 such that $M_1 \subseteq H_1$, $M \setminus M_1 \subseteq H_2$, $M \cap H_1 \cap H_2 = \emptyset$. Since all classifiers whose sets of positivity correspond to 3 dimensional half planes H_1, H_2 are in G_1, and we know that for all such collections of points M and all subsets $M_1 \subseteq M$ we can find a classifier $g_f \in G_1$ such that $g_f(x) = +1$ for all $x \in M_1$ and $g_f(x) = -1$ for all $x \in M \setminus M_1$. Note that for each unique subset $M_1 \subseteq M$, such a g_f will produce a unique sequences $(g(x_1), ..., g(x_n)) \in \{-1, +1\}^n$. There are 2^n unique subsets $M_1 \subseteq M$, so for any M shattered by G, we know

$$|(G_1)_M| := |\{(g(x_1), ..., g(x_n))|g \in G_1\}| = 2^n$$

for all M shattered by G. So G_1 shatters all collections M shattered by G. By the definition of the growth function τ, we know $\tau_G(n) = 2^n \implies \tau_{G_1}(n) = 2^n$. By the definition of VC dimension, we have

$$VC(G) := \max\{n \in \mathbb{N} : \tau_G(n) = 2^n\} \leq \max\{n \in \mathbb{N} : \tau_{G_1}(n) = 2^n\} := VC(G_1)$$

This completes our proof that $VC(G) \leq VC(G_1)$. Since we already showed $VC(G_1) = 4$, we know

$$VC(G) \leq 4$$

Claim: G shatters the set of points $M = \{x_1, x_2, x_3, x_4\} = \{(0,0,0), (0,3,0), (3,0,3), (3,0,-3)\}$.
Proof: It suffices to show that $|G_M| = 2^{|M|} = 2^4 = 16$. Consider spheres of the form

$$S = \{(x, y, z)|(x-a)^2 + (y-b)^2 + (z-c)^2 \leq r^2\}$$

so we can write the classifiers $g_S \in G$ as

$$g_S(x, y, z) := \begin{cases} +1 \text{ if } (x, y, z) \in S \\ -1 \text{ if } (x, y, z) \notin S \end{cases}$$

Consider the spheres

$$\begin{aligned}
S_1 &:= \{(x,y,z)|x^2+(y+2)^2+z^2 \leq 1\} \\
S_2 &:= \{(x,y,z)|x^2+y^2+z^2 \leq 30\} \\
S_3 &:= \{(x,y,z)|x^2+y^2+z^2 \leq 1\} \\
S_4 &:= \{(x,y,z)|x^2+(y-3)^2+z^2 \leq 1\} \\
S_5 &:= \{(x,y,z)|(x-3)^2+y^2+(z-3)^2 \leq 1\} \\
S_6 &:= \{(x,y,z)|(x-3)^2+y^2+(z+3)^2 \leq 1\} \\
S_7 &:= \{(x,y,z)|x^2+y^2+z^2 \leq 10\} \\
S_8 &:= \{(x,y,z)|x^2+y^2+(z+3)^2 \leq 10\} \\
S_9 &:= \{(x,y,z)|(x-10)^2+y^2+z^2 \leq 70\} \\
S_{10} &:= \{(x,y,z)|x^2+y^2+(z-3)^2 \leq 10\} \\
S_{11} &:= \{(x,y,z)|(x-7)^2+(y-5)^2+(z+3)^2 \leq 70\} \\
S_{12} &:= \{(x,y,z)|(x-7)^2+(y-5)^2+(z-3)^2 \leq 70\} \\
S_{13} &:= \{(x,y,z)|x^2+(y+3)^2+z^2 \leq 33\} \\
S_{14} &:= \{(x,y,z)|(x-5)^2+(y-5)^2+z^2 \leq 40\} \\
S_{15} &:= \{(x,y,z)|x^2+y^2+(z-3)^2 \leq 20\} \\
S_{16} &:= \{(x,y,z)|x^2+y^2+(z+3)^2 \leq 20\}
\end{aligned}$$

and the corresponding classifiers $g_{S_1}, ..., g_{S_{16}}$. We can directly compute that

$$(g_{S_1}(x_1), g_{S_1}(x_2), g_{S_1}(x_3), g_{S_1}(x_4)) = (-1,-1,-1,-1)$$

since

$$0^2+(0+2)^2+0^2 = 4 \not\leq 1, \quad 0^2+(3+2)^2+0^2 = 25 \not\leq 1,$$
$$3^2+(0+2)^2+3^2 = 22 \not\leq 1, \quad 3^2+(0+2)^2+(-3)^2 = 22 \not\leq 1$$

and

$$(g_{S_2}(x_1), g_{S_2}(x_2), g_{S_2}(x_3), g_{S_2}(x_4)) = (+1,+1,+1,+1)$$

since

$$0^2+0^2+0^2 = 0 \leq 30, \quad 0^2+(3)^2+0^2 = 9 \leq 30,$$
$$3^2+0^2+3^2 = 18 \leq 30, \quad 3^2+0^2+(-3)^2 = 18 \leq 30$$

and

$$(g_{S_3}(x_1), g_{S_3}(x_2), g_{S_3}(x_3), g_{S_3}(x_4)) = (+1,-1,-1,-1)$$

since

$$0^2+0^2+0^2 = 0 \leq 1, \quad 0^2+(3)^2+0^2 = 9 \not\leq 1,$$
$$3^2+0^2+3^2 = 18 \not\leq 1, \quad 3^2+0^2+(-3)^2 = 18 \not\leq 1$$

and

$$(g_{S_4}(x_1), g_{S_4}(x_2), g_{S_4}(x_3), g_{S_4}(x_4)) = (-1,+1,-1,-1)$$

since

$$0^2+(0-3)^2+0^2 = 9 \not\leq 1, \quad 0^2+(3-3)^2+0^2 = 0 \leq 1,$$
$$3^2+(0-3)^2+3^2 = 27 \not\leq 1, \quad 3^2+(0-3)^2+(-3)^2 = 27 \not\leq 1$$

and

$$(g_{S_5}(x_1), g_{S_5}(x_2), g_{S_5}(x_3), g_{S_5}(x_4)) = (-1, -1, +1, -1)$$

since

$$(0-3)^2+0^2+(0-3)^2 = 18 \not\leq 1, \quad (0-3)^2+(3)^2+(0-3)^2 = 27 \not\leq 1,$$
$$(3-3)^2+0^2+(3-3)^2 = 0 \leq 1, \quad (3-3)^2+0^2+(-3-3)^2 = 36 \not\leq 1$$

and

$$(g_{S_6}(x_1), g_{S_6}(x_2), g_{S_6}(x_3), g_{S_6}(x_4)) = (-1, -1, -1, +1)$$

since

$$(0-3)^2+0^2+(0+3)^2 = 18 \not\leq 1, \quad (0-3)^2+(3)^2+(0+3)^2 = 27 \not\leq 1,$$
$$(3-3)^2+0^2+(3+3)^2 = 36 \not\leq 1, \quad (3-3)^2+0^2+(-3+3)^2 = 0 \leq 1$$

and

$$(g_{S_7}(x_1), g_{S_7}(x_2), g_{S_7}(x_3), g_{S_7}(x_4)) = (+1, +1, -1, -1)$$

since

$$0^2+0^2+0^2 = 0 \leq 10, \quad 0^2+(3)^2+0^2 = 9 \leq 10,$$
$$3^2+0^2+3^2 = 18 \not\leq 10, \quad 3^2+0^2+(-3)^2 = 18 \not\leq 10$$

and

$$(g_{S_8}(x_1), g_{S_8}(x_2), g_{S_8}(x_3), g_{S_8}(x_4)) = (+1, -1, -1, +1)$$

since

$$0^2+0^2+(0+3)^2 = 9 \leq 10, \quad 0^2+(3)^2+(0+3)^2 = 18 \not\leq 10,$$
$$3^2+0^2+(3+3)^2 = 45 \not\leq 10, \quad 3^2+0^2+(-3+3)^2 = 9 \leq 10$$

and

$$(g_{S_9}(x_1), g_{S_9}(x_2), g_{S_9}(x_3), g_{S_9}(x_4)) = (-1, -1, +1, +1)$$

since

$$(0-10)^2+0^2+0^2 = 100 \not\leq 70, \quad (0-10)^2+(3)^2+0^2 = 109 \not\leq 70,$$
$$(3-10)^2+0^2+3^2 = 58 \leq 70, \quad (3-10)^2+0^2+(-3)^2 = 58 \leq 1$$

and

$$(g_{S_{10}}(x_1), g_{S_{10}}(x_2), g_{S_{10}}(x_3), g_{S_{10}}(x_4)) = (+1, -1, +1, -1)$$

since

$$0^2+0^2+(0-3)^2 = 9 \leq 10, \quad 0^2+(3)^2+(0-3)^2 = 18 \not\leq 10,$$
$$3^2+0^2+(3-3)^2 = 9 \leq 10, \quad 3^2+0^2+(-3-3)^2 = 45 \not\leq 10$$

and

$$(g_{S_{11}}(x_1), g_{S_{11}}(x_2), g_{S_{11}}(x_3), g_{S_{11}}(x_4)) = (-1, +1, -1, +1)$$

since

$$(0-7)^2+(0-5)^2+(0+3)^2 = 83 \not\leq 70, \quad (0-7)^2+(3-5)^2+(0+3)^2 = 62 \leq 70,$$
$$(3-7)^2+(0-5)^2+(3+3)^2 = 77 \not\leq 70, \quad (3-7)^2+(0-5)^2+(-3+3)^2 = 41 \leq 70$$

and

$$(g_{S_{12}}(x_1), g_{S_{12}}(x_2), g_{S_{12}}(x_3), g_{S_{12}}(x_4)) = (-1, +1, +1, -1)$$

since

$$(0-7)^2+(0-5)^2+(0-3)^2=83\not\leq 70,\quad (0-7)^2+(3-5)^2+(0-3)^2=62\leq 70,$$
$$(3-7)^2+(0-5)^2+(3-3)^2=41\leq 70,\quad (3-7)^2+(0-5)^2+(-3-3)^2=77\not\leq 70$$

and

$$(g_{S_{13}}(x_1),g_{S_{13}}(x_2),g_{S_{13}}(x_3),g_{S_{13}}(x_4))=(+1,-1,+1,+1)$$

since

$$0^2+(0+3)^2+0^2=9\leq 33,\quad 0^2+(3+3)^2+0^2=36\not\leq 33,$$
$$3^2+(0+3)^2+3^2=27\leq 33,\quad 3^2+(0+3)^2+(-3)^2=27\leq 33$$

and

$$(g_{S_{14}}(x_1),g_{S_{14}}(x_2),g_{S_{14}}(x_3),g_{S_{14}}(x_4))=(-1,+1,+1,+1)$$

since

$$(0-5)^2+(0-5)^2+0^2=50\not\leq 40,\quad (0-5)^2+(3-5)^2+0^2=29\leq 40,$$
$$(3-5)^2+(0-5)^2+3^2=38\leq 40,\quad (3-5)^2+(0-5)^2+(-3)^2=38\leq 40$$

and

$$(g_{S_{15}}(x_1),g_{S_{15}}(x_2),g_{S_{15}}(x_3),g_{S_{15}}(x_4))=(+1,+1,+1,-1)$$

since

$$0^2+0^2+(0-3)^2=9\leq 20,\quad 0^2+(3)^2+(0-3)^2=18\leq 20,$$
$$3^2+0^2+(3-3)^2=9\leq 20,\quad 3^2+0^2+(-3-3)^2=45\not\leq 20$$

and

$$(g_{S_{16}}(x_1),g_{S_{16}}(x_2),g_{S_{16}}(x_3),g_{S_{16}}(x_4))=(+1,+1,-1,+1)$$

since

$$0^2+0^2+(0+3)^2=9\leq 10,\quad 0^2+(3)^2+(0+3)^2=18\leq 20,$$
$$3^2+0^2+(3+3)^2=45\not\leq 20,\quad 3^2+0^2+(-3+3)^2=9\leq 20$$

Thus,

$$(g_i(x_1),g_i(x_2),g_i(x_3),g_i(x_4))\neq(g_j(x_1),g_j(x_2),g_j(x_3),g_j(x_4))$$

for all $i,j\in\{1,...,16\}$ such that $i\neq j$. Since there are 16 such distinct sequences $(g_i(x_1),g_i(x_2),g_i(x_3),g_i(x_4))$, we know

$$|G_M|:=|\{(g(x_1),g(x_2),g(x_3),g(x_4))|g\in G\}|=16=2^4=2^{|M|}$$

This completes the proof that G shatters M. Since $|M|=4$, we know $\tau_G(4)=2^4=16$. By the definition of VC dimension, we have

$$VC(G):=\max\{n\in\mathbb{N}:\tau_G(n)=2^n\}\geq 4$$

since $4\in\{n\in\mathbb{N}:\tau_G(n)=2^n\}$. However, we already proved that

$$VC(G)\leq VC(G_1)=4$$

so we can conclude that

$$VC(G)=4$$

This completes the proof that the VC dimension of the class of classifiers that take value +1 inside spheres in 3 dimensional space is $VC(G)=4$, which completes part (d).

5. Let F be the set of all polynomials of degree at most 3 in 2 variables x and y, that is, $f(x,y) = \sum \alpha_j x^{a_j} y^{b_j}$ where a_j and b_j are non-negative integers such that $a_j + b_j \leq 3$. Let G be the set of binary classifiers such that every $g \in G$ takes value $+1$ on a set $f(x,y) > 0$ for some $f \in F$. Find the VC dimension of G.

Solution.
Claim: The VC dimension of G is $VC(G) = 10$.
Proof: Note that the combinations of exponents a_j, b_j such that $a_j, b_j \geq 0$, $a_j + b_j \leq 3$ are fixed with

$$(a_j, b_j) \in \{(0,0),(1,0),(0,1),(2,0),(1,1),(0,2),(3,0),(2,1),(1,2),(0,3)\}$$

Since $|\{(0,0),(1,0),(0,1),(2,0),(1,1),(0,2),(3,0),(2,1),(1,2),(0,3)\}| = 10$, we can let

$$\begin{aligned}\{(a_j,b_j)|a_j,b_j \geq 0, a_j+b_j \leq 3\} = \quad & \{(a_1,b_1),(a_2,b_2),(a_3,b_3),(a_4,b_4),(a_5,b_5),\\ & (a_6,b_6),(a_7,b_7),(a_8,b_8),(a_9,b_9),(a_{10},b_{10})\}\\ = \quad & \{(0,0),(1,0),(0,1),(2,0),(1,1),\\ & (0,2),(3,0),(2,1),(1,2),(0,3)\}\end{aligned}$$

Then we can write two arbitrary functions $f_1, f_2 \in F$ as

$$f_1(x,y) := \sum_{j=1}^{10} c_j x^{a_j} y^{b_j} \quad f_2(x,y) := \sum_{j=1}^{10} d_j x^{a_j} y^{b_j}$$

where $c_1, ..., c_{10}, d_1, ..., d_{10} \in \mathbb{R}$ are scalars. Then for all scalars $u, v \in \mathbb{R}$, we have

$$\begin{aligned} uf_1(x,y) + vf_2(x,y) = \quad & u\sum_{j=1}^{10} c_j x^{a_j} y^{b_j} + v\sum_{j=1}^{10} d_j x^{a_j} y^{b_j}\\ = \quad & \sum_{j=1}^{10} uc_j x^{a_j} y^{b_j} + \sum_{j=1}^{10} vd_j x^{a_j} y^{b_j}\\ = \quad & \sum_{j=1}^{10} (uc_j + vd_j) x^{a_j} y^{b_j}\end{aligned}$$

Note that $u, v, c_1, ..., c_{10}, d_1, ..., d_{10} \in R$ implies $uc_J + vd_j \in \mathbb{R}$ for all $j \in \{1, ..., 10\}$, so we can let $\alpha_j = uc_j + vd_j$ for all $j \in \{1, ..., 10\}$ to find

$$uf_1(x,y) + vf_2(x,y) = \sum_{j=1}^{10} \alpha_j x^{a_j} y^{b_j} = \sum \alpha_j x^{a_j} y^{b_j} = f_3(x,y) \in F$$

Thus, F is closed under addition and scalar multiplication, so we know F is a linear space. Since the problem statement defines G to be

$$G := \{g_f | f \in F\}$$

where

$$g_f(x,y) = \begin{cases} +1 \text{ if } f(x,y) > 0 \\ -1 \text{ if } f(x,y) < 0 \end{cases}$$

the R. Dudley Theorem from lecture guarantees that if F has finite dimension $dim(F) < \infty$, then $VC(G) = dim(F)$. Thus, we just need to prove that F has finite dimension and compute $dim(F)$. Note that, for all $f \in F$, f takes the form

$$f(x,y) = \alpha_1(1) + \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_5 xy + \alpha_6 y^2 + \alpha_7 x^3 + \alpha_8 x^2 y + \alpha_9 xy^2 + \alpha_{10} y^3$$

Thus, all $f \in F$ can be written as a linear combination of the vectors in

$$B := \{(1), (x), (y), (x^2), (xy), (y^2), (x^3), (x^2y), (xy^2), (y^3)\}$$

so B spans F. Furthermore, none of the vectors in B can be expressed as a linear combination of each other. That is,

$$1 \neq \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_5 xy + \alpha_6 y^2 + \alpha_7 x^3 + \alpha_8 x^2 y + \alpha_9 xy^2 + \alpha_{10} y^3$$
$$\forall \alpha_2, ..., \alpha_{10} \in \mathbb{R}$$
$$x \neq \alpha_1(1) + \alpha_3 y + \alpha_4 x^2 + \alpha_5 xy + \alpha_6 y^2 + \alpha_7 x^3 + \alpha_8 x^2 y + \alpha_9 xy^2 + \alpha_{10} y^3$$
$$\forall \alpha_1, \alpha_3, ..., \alpha_{10} \in \mathbb{R}$$
$$y \neq \alpha_1(1) + \alpha_2 x + \alpha_4 x^2 + \alpha_5 xy + \alpha_6 y^2 + \alpha_7 x^3 + \alpha_8 x^2 y + \alpha_9 xy^2 + \alpha_{10} y^3$$
$$\forall \alpha_1, \alpha_2, \alpha_4, ..., \alpha_{10} \in \mathbb{R}$$
$$x^2 \neq \alpha_1(1) + \alpha_2 x + \alpha_3 y + \alpha_5 xy + \alpha_6 y^2 + \alpha_7 x^3 + \alpha_8 x^2 y + \alpha_9 xy^2 + \alpha_{10} y^3$$
$$\forall \alpha_1, ..., \alpha_3, \alpha_5, ..., \alpha_{10} \in \mathbb{R}$$
$$xy \neq \alpha_1(1) + \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_6 y^2 + \alpha_7 x^3 + \alpha_8 x^2 y + \alpha_9 xy^2 + \alpha_{10} y^3$$
$$\forall \alpha_1, ..., \alpha_4, \alpha_6, ..., \alpha_{10} \in \mathbb{R}$$
$$y^2 \neq \alpha_1(1) + \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_5 xy + \alpha_7 x^3 + \alpha_8 x^2 y + \alpha_9 xy^2 + \alpha_{10} y^3$$
$$\forall \alpha_1, ..., \alpha_5, \alpha_5, \alpha_7, ..., \alpha_{10} \in \mathbb{R}$$
$$x^3 \neq \alpha_1(1) + \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_5 xy + \alpha_6 y^2 + \alpha_8 x^2 y + \alpha_9 xy^2 + \alpha_{10} y^3$$
$$\forall \alpha_1, ..., \alpha_6, \alpha_8, ..., \alpha_{10} \in \mathbb{R}$$
$$x^2 y \neq \alpha_1(1) + \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_5 xy + \alpha_6 y^2 + \alpha_7 x^3 + \alpha_9 xy^2 + \alpha_{10} y^3$$
$$\forall \alpha_1, ..., \alpha_7, \alpha_9, \alpha_{10} \in \mathbb{R}$$
$$xy^2 \neq \alpha_1(1) + \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_5 xy + \alpha_6 y^2 + \alpha_7 x^3 + \alpha_8 x^2 y + \alpha_{10} y^3$$
$$\forall \alpha_1, ..., \alpha_8, \alpha_{10} \in \mathbb{R}$$
$$y^3 \neq \alpha_1(1) + \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_5 xy + \alpha_6 y^2 + \alpha_7 x^3 + \alpha_8 x^2 y + \alpha_9 xy^2$$
$$\forall \alpha_1, ..., \alpha_9 \in \mathbb{R}$$

Thus, B is a linearly independent spanning set for F. By the definition of the basis of a linear space, we know B is a basis for F. By the definition of the dimension of linear space, we know

$$dim(L) = |B| = |\{(1), (x), (y), (x^2), (xy), (y^2), (x^3), (x^2y), (xy^2), (y^3)\}| = 10 < \infty$$

Since $dim(L) = 10 < \infty$ is finite, we can apply the R. Dudley Theorem from lecture to find

$$VC(G) = dim(L) = 10$$

This completes the proof that the VC dimension of the set G of binary classifiers that take value $+1$ only when some polynomial $f(x, y)$ of degree at most 3 is positive is $VC(G) = 10$.

Assignment 4

Read Chapter 9 of the textbook "Understanding Machine Learning." Then do the following problems:

1. The Margin:

recall that in the Perceptron algorithm, we assume that the data $(X_1, Y_1), ..., (X_n, Y_n)$ are such that $(X_j, Y_j) \in \mathbb{R}^d \times \{\pm 1\}$ and that there exists a vector $w \in \mathbb{R}^d$ of unit length such that $(Y_j < w, X_j >) > 0$. The *margin* of a hyperplane $H_w = \{x :< w, x >= 0\}$ is the smallest among the distances from $X_1, ..., X_n$ to H_w.

(a) Show that the margin equals $min_{j\in\{1,...,n\}}| < w, X_j > |$.
(hint: recall that the vector w is orthogonal to the hyperplane H_w. Now draw a picture in $2d$ and convince yourself, and then me, that $| < w, X_j > |$ is the length of the projection of X_j onto w.)

(b) Given two hyperplanes H_{w_1} and H_{w_2} that both separate the classes labeled $+1$ and -1, which one would be a better pick?

Solution.

(a) We want to show that the *margin* of a hyperplane H_w equals

$$min_{j\in\{1,...,n\}}| < w, X_j > |$$

By definition, the margin of H_w is the smallest among the distances from $X_1, ..., X_n$ to H_w. Thus, we are really trying to prove the statement

$$min_{j\in\{1,...,n\}}(\text{distance from } X_j \text{ to } H_w) = min_{j\in\{1,...,n\}}| < w, X_j > | \qquad (1)$$

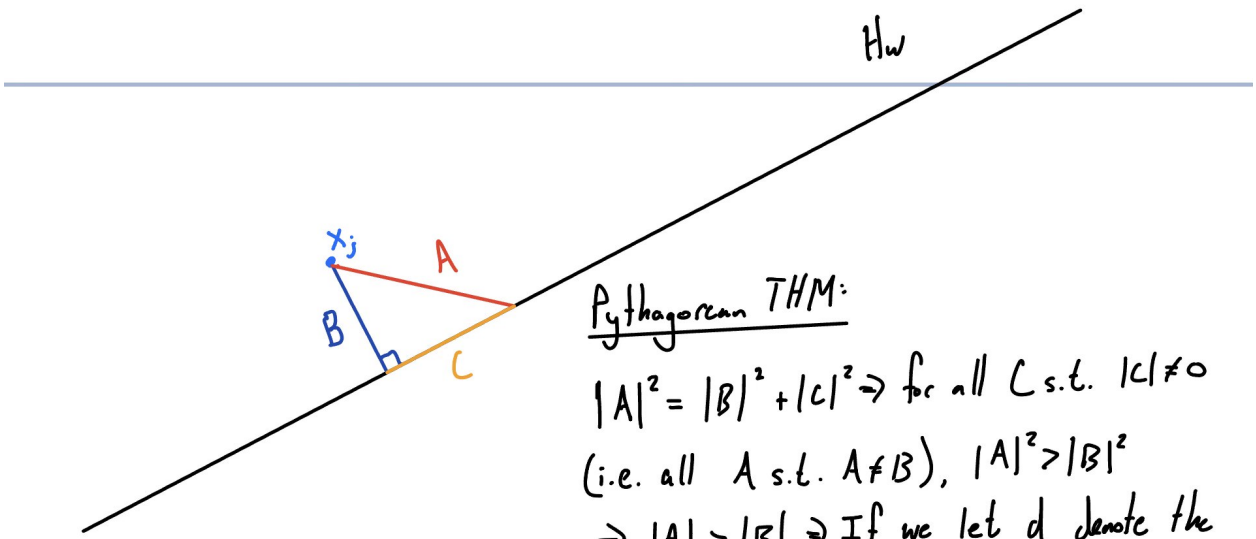
Note that

$$(\text{distance from } X_j \text{ to } H_w) = | < w, X_j > | \quad \text{for all } j \in \{1, ..., n\} \qquad (2)$$

implies (1), so it suffices to show that (2) holds for all $j \in \{1, ..., n\}$. To do so, we will show that the length of the projection of X_j onto w equals

$$| < w, X_j > | \quad \text{for all } j \in \{1, ..., n\} \qquad (3)$$

Note that the distance from X_j to H_w equals the length of the shortest segment connecting X_j to H_w which equals the length of the segment B connecting the X_j to H_w that intersects H_w *orthogonally*. This follows from the Pythagorean theorem, as any segment A from X_j to H_w that doesn't intersect H_w orthogonally is the hypotenuse of a right triangle with sides A, B and the segment C of H_w that connects A and B. A visual depiction helps to shed light on this result:



Pythagorean THM:

$|A|^2 = |B|^2 + |C|^2 \Rightarrow$ for all C s.t. $|C| \neq 0$ (i.e. all A s.t. $A \neq B$), $|A|^2 > |B|^2$

$\Rightarrow |A| > |B| \Rightarrow$ If we let d denote the distance from X_j to H_w, then

$d = |B|$

Thus, we need show $|B|$ (the length of the segment connecting X_j to H_w that intersects H_w orthogonally) equals

$$| < w, X_j > | \quad \text{for all } j \in \{1, ..., n\} \tag{4}$$

Since B intersects H_w orthogonally, and w also intersects H_w orthogonally, we know $B \parallel w$ (B and w are *parallel*). If we consider X_j and w as d dimensional vectors from the origin, then they intersect at the origin. By the definition of B, X_j intersects B. Let a^* be the angle between X_j and w and a_2 be the acute angle between X_j and B. Define

$$a_1 := \begin{cases} a^* \text{ if } 0 \leq a^* \leq \pi \\ \pi - a^* \text{ otherwise} \end{cases}$$

Note that a_1 and a_2 are on opposite sides of the transversal X_j between w and B (if $0 \leq a^* \leq \pi$, then a_1 is above the vector X_j while a_2 is below X_j, whereas if $\pi < a^* < 2\pi$, then a_1 is below the vector X_j while a_2 is above it). Thus, since w and B are parallel, we know a_1 and a_2 are alternate interior angles between parallel lines, so

$$a_1 = a_2 = \theta \tag{5}$$

Also, since a_2 is defined to be acute, (5) implies that both a_1 and a_2 are acute. Thus, a_1 and a_2 can be viewed as the *acute* angles between X_j and w and X_j and B, respectively.
By definition, the projection $proj_v u$ of u onto v is the vector parallel to v with magnitude equal to the distance u covers in the direction of v. Thus, the length of the projection of u onto v is just the distance u covers in the direction of v. By definition, this magnitude $||proj_v u||$ is equal to $|comp_v u|$, the absolute value of the component of u onto v. For X_j and w, this implies that $|comp_w X_j|$ (the length of the projection of X_j onto w) is just the distance X_j covers in the direction of w. Since $a_1 = \theta$ is always acute, the definition of the cosine function yields

$$cos(a_1) = cos(\theta) = \frac{||proj_w X_j||}{||X_j||} = \frac{|comp_w X_j|}{||X_j||} \tag{6}$$

so for all $a_1 = \theta$, we have

$$|comp_w X_j| = ||proj_w X_j|| = ||X_j|| cos(\theta) = ||X_j|| cos(a_1) \tag{7}$$

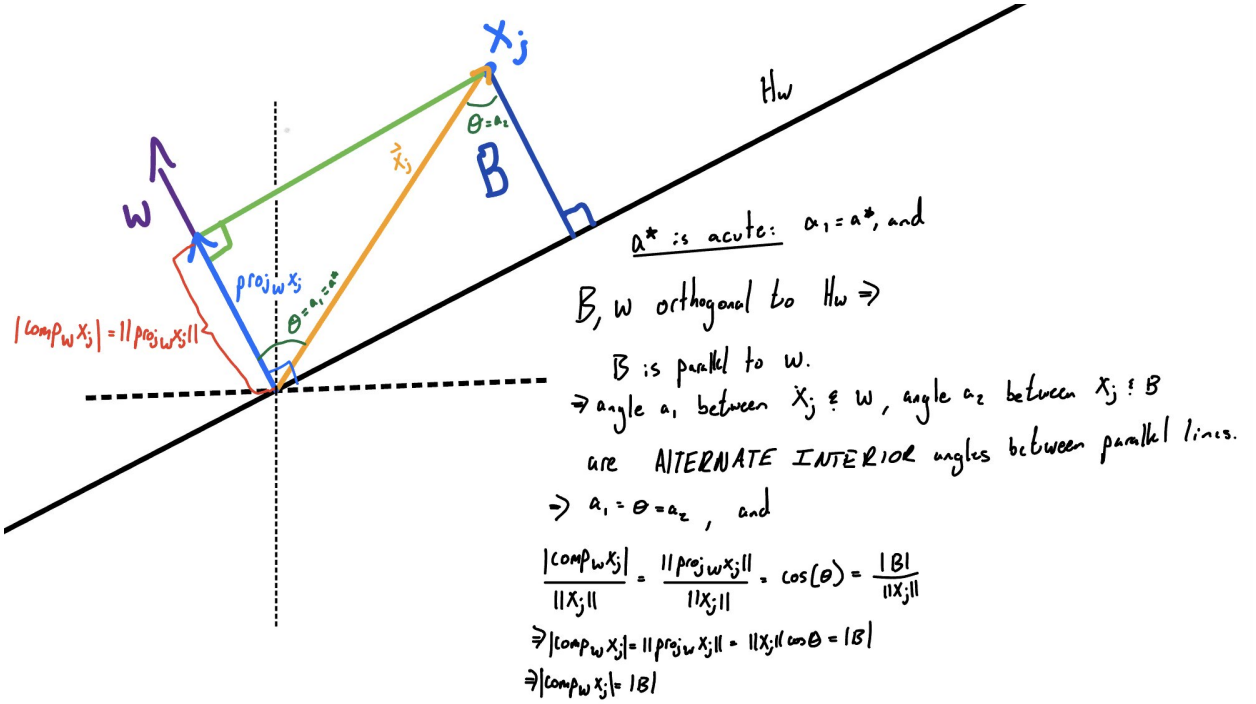
Similarly, since the angle a_2 is always acute, the definition of the cosine function yields

$$cos(a_2) = cos(\theta) = \frac{|B|}{||X_j||} \implies |B| = ||X_j|| cos(\theta) \tag{8}$$

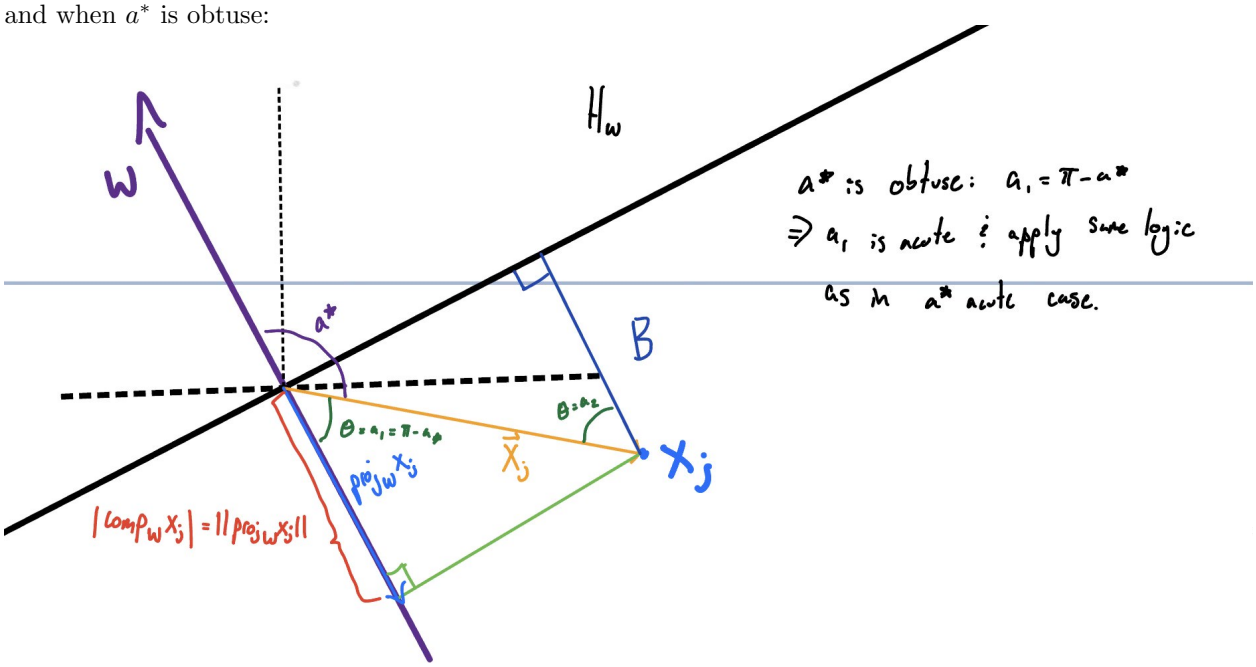
Comparing (7) and the right-most equation from (8), we see that for all acute angles $\theta < \pi$ between X_j and w and X_j and B,

$$||proj_w X_j|| = |comp_w X_j| = |B| \tag{9}$$

We provide a visual depiction when a^* is acute:



and when a^* is obtuse:



to clarify the preceding argument.
Combining (9) with (4), we see it suffices to show that the length of the projection of X_j onto w satisfies

$$|comp_w X_j| = ||proj_w X_j|| = | < w, X_j > | \text{ for all } j \in \{1, ..., n\} \qquad (10)$$

Note that (10) is formalizing the sufficient condition we alluded to in (3). By definition of the dot product, we have

$$< w, X_j >= ||w|| \cdot ||X_j|| \cdot cos(a^*) \qquad (11)$$

and the cosine function satisfies

$$cos(a^*) = -cos(\pi - a^*) \qquad (12)$$

Thus, if a^* is obtuse, we have

$$| < w, X_j > | = ||w|| \cdot ||X_j|| \cdot |cos(a^*)| = ||w|| \cdot ||X_j|| \cdot | - cos(a_1)| = ||w|| \cdot ||X_j|| \cdot cos(a_1) \qquad (13)$$

(since $cos(a_1) \geq 0$ for all acute a_1), and if a^* is acute, we have

$$| < w, X_j > | = ||w|| \cdot ||X_j|| \cdot |cos(a^*)| = ||w|| \cdot ||X_j|| \cdot |cos(a_1)| = ||w|| \cdot ||X_j|| \cdot cos(a_1) \qquad (14)$$

Comparing (13) and (14) with (7), we see that regardless of whether a^* is acute or obtuse, we have

$$| < w, X_j > | = ||w|| \cdot |comp_w X_j| = ||proj_w X_j|| \quad \text{for all } j \in \{1, ..., n\} \qquad (15)$$

However, we are given that w has unit length in the problem statement, so we know

$$| < w, X_j > | = 1 \cdot |comp_w X_j| = |comp_w X_j| = ||proj_w X_j|| \quad \text{for all } j \in \{1, ..., n\} \qquad (16)$$

That is, the absolute value of the dot product of w and X_j equals the length of the projection of X_j onto w for all $j \in \{1, ..., n\}$. Since $|B|$ is the distance from X_j to H_w for all $j \in \{1, ..., n\}$, we know from (9) that

$$(\text{ distance from } X_j \text{ to } H_w) = |B| = |comp_w X_j| = ||proj_w X_j|| = | < w, X_j > | \quad \text{for all } j \in \{1, ..., n\} \qquad (17)$$

Since (17) holds for all $j \in \{1, ..., n\}$, it implies

$$min_{j \in \{1,...,n\}}(\text{ distance from } X_j \text{ to } H_w) = \min_{j \in \{1,...,n\}} | < w, X_j > | \qquad (18)$$

By the definition of the margin of a hyperplane H_w, (18) completes the proof that the margin equals $min_{j \in \{1,...,n\}} | < w, X_j > |$.

(b) *Claim:* If the margin of H_{w_1} does not equal the margin of H_{w_2}, we should choose the hyperplane with the larger margin.
Proof. Denote

$$M_1 := \min_{j \in \{1,..n\}} | < w_1, X_j > | \quad M_2 := \min_{j \in \{1,...,n\}} | < w_2, X_j > | \qquad (19)$$

By part (a), M_1 is the margin of H_{w_1} and M_2 is the margin of H_{w_2}. Since our claim only addresses the situation when $M_1 \neq M_2$, we can assume that $M_1 \neq M_2$. Note that, since H_{w_1} and H_{w_2} both separate the classes labeled $+1$ and -1, we know that

$$(Y_j < w_1, X_j >), (Y_j < w_2, X_j >) > 0 \qquad (20)$$

However, this does not rule out the possibility that $\exists$ an instance $X \in \mathbb{R}^d$ and a corresponding label $Y \in \{\pm 1\}$ such that

$$(Y < w_1, X >) \leq 0 \text{ and/or } (Y < w_2, X >) \leq 0 \qquad (21)$$

The left-most inequality from (21) would imply that H_{w_1} misclassifies X while the right-most inequality would imply that H_{w_2} misclassifies X. Since H_{w_1} and H_{w_2} could both still potentially misclassify instances, the 'better pick' will be the hyperplane that minimizes the chances of misclassifying instances $X \in R^d$ (i.e. the most generalizable hyperplane). Note that, for any hyperplane $H_w := \{x \in R^d :< w, x >= 0\}$, as an instance $X \in R^d$ approaches H_w, i.e. as

$$| < w, X > | \to 0$$

then, for the corresponding label $Y \in \{\pm 1\}$, we have

$$Y < w, X > \to 0 \qquad (22)$$

That is, as instances approach any hyperplane, that hyperplane gets closer to misclassifying them. Although instances could potentially vary from training instances enough to cause H_{w_1} and/or H_{w_2} to misclassify them, we still expect a given instance with label $+1$ to be closer to the training instances labeled $+1$ than training instances labeled -1, and vice versa for a given instance with label -1. Thus, to minimize the chance of misclassifying instances, we should pick the hyperplane that maximizes the closest distance from some training instance $X \in \{X_1, ..., X_n\}$ to the hyperplane. Since the distance from X_j to H_w is $| < w, X_j > |$ for all $j \in \{1, ..., n\}$, this is equivalent to maximizing

$$min_{j \in \{1,...,n\}} | < w, X_j > | \qquad (23)$$

Since we proved in part (a) that the margin M_w of a hyperplane H_w is

$$M_W = \min_{j \in \{1,...,n\}} | < w, X_j > |$$

(23) implies that we can minimize the chance of misclassifying instances by choosing the hyperplane with the larger margin. That is, if $M_1 > M_2$, we should choose H_{w_1}, and if $M_2 > M_1$, we should choose H_{w_2}. This completes the proof that the 'better pick' would be the hyperplane with the larger margin.

Note: If $M_1 = M_2$, then the margin provides no information about which hyperplane is a better pick. In this case, we would need to examine different properties of H_{w_1} and H_{w_2} to determine which hyperplane is more generalizable. However, as our sample size n increases, the probability that $M_1 = M_2$ decreases (assuming $w_1 \neq cw_2$ for any scalar $c \in \mathbb{R}$), so the margin will be a useful indicator of the 'better pick' in most cases. Also if $w_1 = cw_2$ for any $c \in \mathbb{R}$, then $H_{w_1} = H_{w_2}$, so the choice of hyperplane would make no difference in generalizability.

2. Perceptron Algorithm: an example.

Consider the following set of labeled points in $\mathbb{R}^2$:

$$\{(1,0), +1\}, \{(2,0), +1\}, \{(0,1), -1\}, \{(1,1), -1\}$$

(a) Can the points be labeled $+1$ and -1 be separated by the hyperplane?

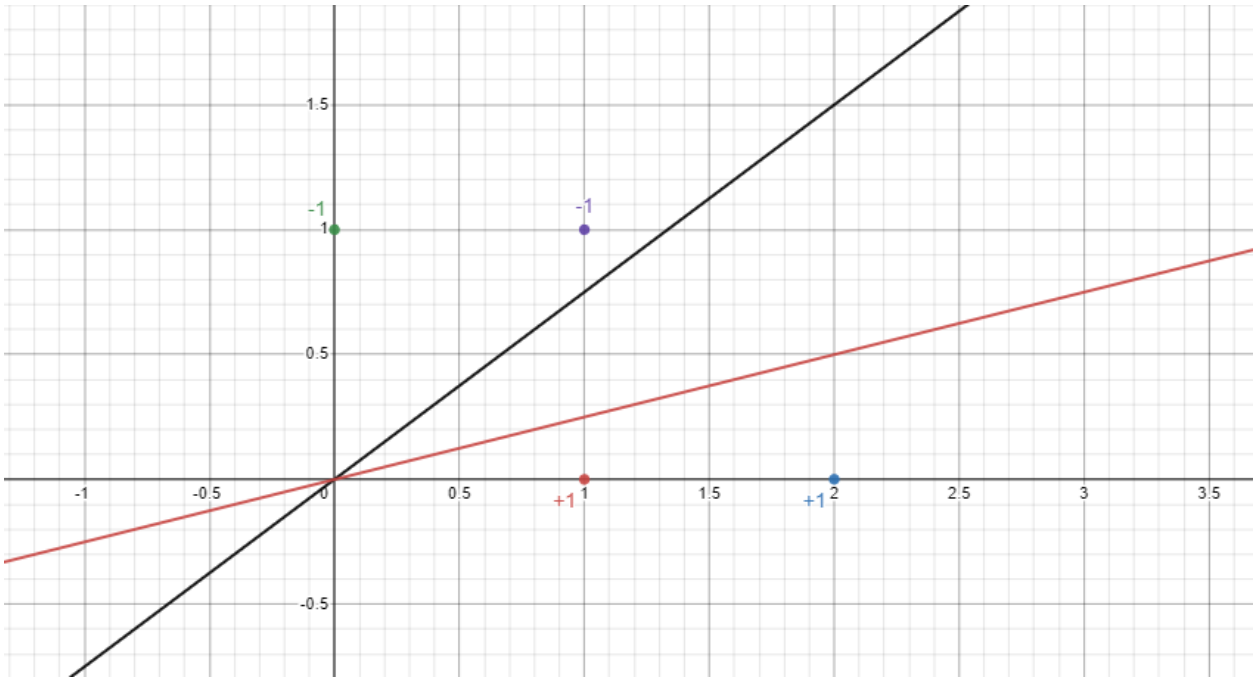
(b) Perform as many iterations of the Perceptron algorithm as necessary to get a separating hyperplane; write the updates explicitly.

(c) What is the margin of the resulting hyperplane?

(d) Based on the number of iterations you needed for convergence and the theorem studied in class on the maximum number of iterations, find an upper bound for the maximum achievable margin.
(recall that we showed in class that the maximal number of iterations T satisfies $T \leq (\frac{R}{\gamma})^2$ where R is the largest norm of X_j's and $\gamma = minY_j < w^*, X_j >$ where $\|w^*\|_2 = 1$)

Solution.

(a) After plotting the points with their corresponding labels, we easily see that they *can* be separated by the hyperplane. To demonstrate this, we plot the labeled points and also plot the hyperplanes $H_1 = \{x \in R^2 :< w_1, x >= 0\}$ and $H_2 = \{x \in \mathbb{R}^2 :< w_2, x >= 0\}$, where $w_1 = (-\frac{1}{4}, 1)$ and $w_2 = (-\frac{3}{4}, 1)$:



Clearly, both hyperplanes separate the points labeled -1 from those labeled $+1$. Thus, the points *can* be separated by the hyperplane.

(b) Before the iterative loop starts, we have the training data

$$(X_1, Y_1), ..., (X_n, Y_n) = \{(1,0), +1\}, \{(2,0), +1\}, \{(0,1), -1\}, \{(1,1), -1\} \qquad (24)$$

and the prediction vector

$$w^{(1)} = (0,0) \qquad (25)$$

We now begin iterating: *Iteration 1:* We have

$$Y_1 < w^{(1)}, X_1 >= 1 < (0,0), (1,0) >= 1(0(1) + 0(0)) = 0 \leq 0$$

so we proceed to the second iteration with

$$w^{(2)} = w^{(1)} + Y_1 X_1 = (0,0) + 1(1,0) = (1,0)$$

Iteration 2: We have

$$\begin{aligned} Y_1 < w^{(2)}, X_1 > = & \quad 1 < (1,0), (1,0) >= 1(1(1) + 0(0)) = 1 \not\leq 0 \\ Y_2 < w^{(2)}, X_2 > = & \quad 1 < (1,0), (2,0) >= 1(1(2) + 0(0)) = 2 \not\leq 0 \\ Y_3 < w^{(2)}, X_3 > = & \quad -1 < (1,0), (0,1) >= -1(1(0) + 0(1)) = 0 \leq 0 \end{aligned}$$

so we proceed to the third iteration with

$$w^{(3)} = w^{(2)} + Y_3 X_3 = (1,0) + -1(0,1) = (1,-1)$$

Iteration 3: We have

$$\begin{aligned} Y_1 < w^{(3)}, X_1 > = & \quad 1 < (1,-1), (1,0) >= 1(1(1) + -1(0)) = 1 \not\leq 0 \\ Y_2 < w^{(3)}, X_2 > = & \quad 1 < (1,-1), (2,0) >= 1(1(2) + -1(0)) = 2 \not\leq 0 \\ Y_3 < w^{(3)}, X_3 > = & \quad -1 < (1,-1), (0,1) >= -1(1(0) + -1(1)) = -(-1) = 1 \not\leq 0 \\ Y_4 < w^{(3)}, X_4 > = & \quad -1 < (1,-1), (1,1) >= -1(1(1) + -1(1)) = -1(0) = 0 \leq 0 \end{aligned}$$

so we proceed to the fourth iteration with

$$w^{(4)} = w^{(3)} + Y_4 X_4 = (1,-1) + -1(1,1) = (0,-2)$$

Iteration 4: We have

$$Y_1 < w^{(4)}, X_1 >= 1((0,-2),(1,0)) = 1(0(1) + -2(0)) = 0 \leq 0$$

so we proceed to the fifth iteration with

$$w^{(5)} = w^{(4)} + Y_1 X_1 = (0,-2) + 1(1,0) = (1,-2)$$

Iteration 5: Now, we find

$$\begin{aligned}
Y_1 < w^{(5)}, X_1 >= &\quad 1 < (1,-2),(1,0) >= 1(1(1) + -2(0)) = 1 \not\leq 0 \\
Y_2 < w^{(5)}, X_2 >= &\quad 1 < (1,-2),(2,0) >= 1(1(2) + -2(0)) = 2 \not\leq 0 \\
Y_3 < w^{(5)}, X_3 >= &\quad -1 < (1,-2),(0,1) >= -1(1(0) + -2(1)) = -(-2) = 2 \not\leq 0 \\
Y_4 < w^{(5)}, X_4 >= &\quad -1 < (1,-2),(1,1) >= -1(1(1) + -2(1)) = -1(1-2) = -(-1) = 1 \not\leq 0
\end{aligned}$$

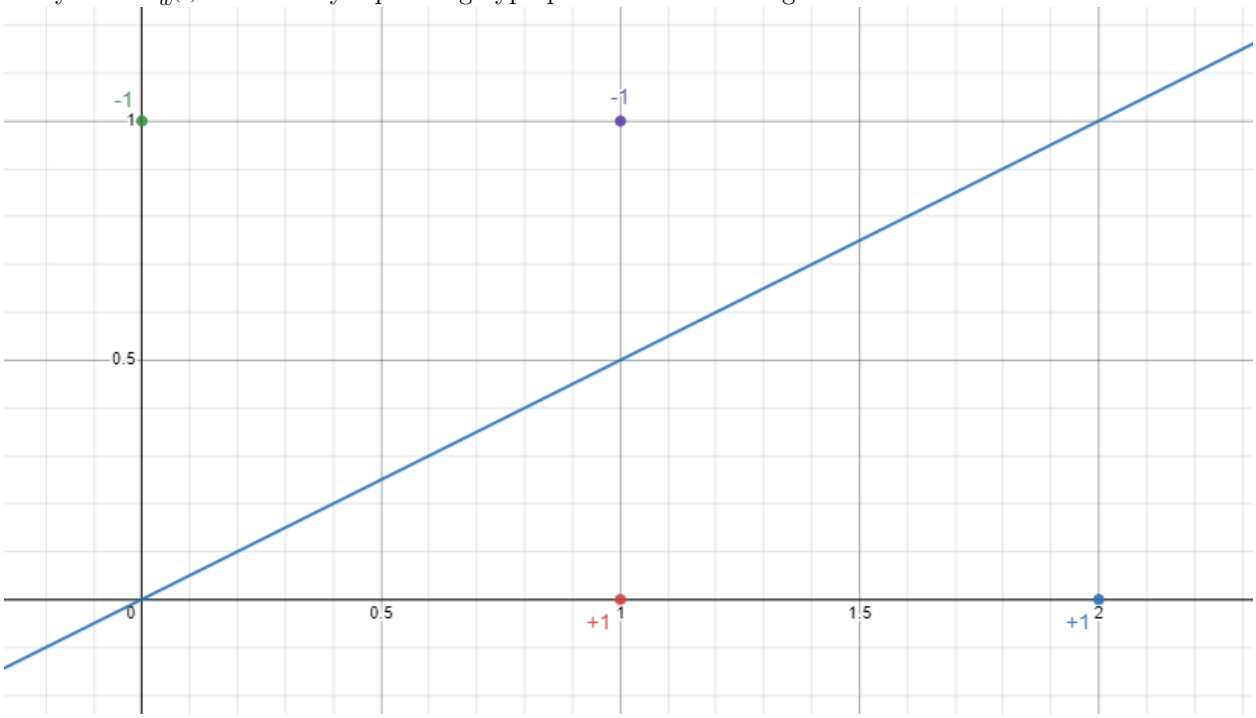
so the Perceptron Algorithm converges and outputs $w^{(5)} = (1,-2)$ as the prediction vector. From this vector, we construct the hyperplane

$$H_{w^{(5)}} = \{x \in R^2 :< w^{(5)}, x >= 0\}$$

Note that $H_{w^{(5)}}$ includes all points $(x,y) \in \mathbb{R}^2$ such that

$$< (1,-2),(x,y) >= x - 2y = 0 \iff y = \frac{x}{2}$$

so the hyperplane returned by this execution of the Perceptron algorithm is just the line $y = \frac{x}{2}$. Note that the hyperplane $H_{w^{(5)}}$ classifies all points below $y = \frac{x}{2}$ with label $+1$ and all points above $y = \frac{x}{2}$ with label -1. By plotting $H_{w^{(5)}}$ alongside the labeled points from our training data, we can easily verify that $H_{w^{(5)}}$ is a linearly separating hyperplane for the training data:



which completes part (b).

(c) From part (a) of the first exercise, we know that the margin M_{H_w} of any hyperplane $H_w := \{x :< w, x >= 0\}$ for any vector w of magnitude $||w|| = 1$ is

$$M_{H_w} := \min_{j \in \{1,...,n\}} (\text{ distance from } X_j \text{ to } H_w) = \min_{j \in \{1,...,n\}} | < w, X_j > | \qquad (26)$$

The vector returned by our execution of the Perceptron algorithm has magnitude

$$||w^{(5)}|| = \sqrt{1^2 + (-2)^2} = \sqrt{1+4} = \sqrt{5} \qquad (27)$$

Fortunately, note that

$$\{x :< w^{(5)}, x >= 0\} = \{x :< cw^{(5)}, x >= 0\} \qquad (28)$$

for all scalars $c \in \mathbb{R}$ s.t. $c \neq 0$. Thus, we can let $c = \frac{1}{||w^{(5)}||}$ and redefine the separating hyperplane identified by our execution of the Perceptron algorithm as

$$H_{w^{(5)}} = \{x :< w^{(5)}, x >= 0\} = \{x :< \frac{w^{(5)}}{||w^{(5)}||}, x >= 0\} = \{x :< w, x >= 0\} = H_w \qquad (29)$$

where

$$w = cw^{(5)} = \frac{w^{(5)}}{||w^{(5)}||} = \frac{1}{\sqrt{5}}(1, -2) = (\frac{1}{\sqrt{5}}, -\frac{2}{\sqrt{5}}) \qquad (30)$$

and we note that

$$||w|| = \sqrt{\frac{1}{\sqrt{5}^2} + (-\frac{2}{\sqrt{5}})^2} = \sqrt{\frac{1}{5} + \frac{4}{5}} = \sqrt{1} = 1 \qquad (31)$$

Thus, we can apply (26) to find that the margin $M_{H_{w^{(5)}}}$ of $H_{w^{(5)}}$, which equals the margin M_{H_w} of H_w, is

$$M_{H_{w^{(5)}}} = M_{H_w} = \min_{j \in \{1,...,n\}} | < w, X_j > | \qquad (32)$$

We can directly compute that

$$\begin{aligned}
| < w, X_1 > | = \quad & | < (\frac{1}{\sqrt{5}}, -\frac{2}{\sqrt{5}}), (1, 0) > | = |\frac{1}{\sqrt{5}}(1) + -\frac{2}{\sqrt{5}}(0)| = \frac{1}{\sqrt{5}} \\
| < w, X_2 > | = \quad & | < (\frac{1}{\sqrt{5}}, -\frac{2}{\sqrt{5}}), (2, 0) > | = |\frac{1}{\sqrt{5}}(2) + -\frac{2}{\sqrt{5}}(0)| = \frac{2}{\sqrt{5}} \\
| < w, X_3 > | = \quad & | < (\frac{1}{\sqrt{5}}, -\frac{2}{\sqrt{5}}), (0, 1) > | = |\frac{1}{\sqrt{5}}(0) + -\frac{2}{\sqrt{5}}(1)| = \frac{2}{\sqrt{5}} \\
| < w, X_4 > | = \quad & | < (\frac{1}{\sqrt{5}}, -\frac{2}{\sqrt{5}}), (1, 1) > | = |\frac{1}{\sqrt{5}}(1) + -\frac{2}{\sqrt{5}}(1)| = \frac{1}{\sqrt{5}}
\end{aligned}$$

so we see that X_1 and X_4 are the closest training instances to the hyperplane $H_{w^{(5)}} = H_w$, both at a distance of $\frac{1}{\sqrt{5}}$. Plugging this result into (32) yields that the margin of $H_{w^{(5)}} = H_w$ is

$$M_{H_{w^{(5)}}} = M_{H_w} = \frac{1}{\sqrt{5}} \qquad (33)$$

Thus, the margin of the hyperplane $H_{w^{(5)}} = H_w$ returned by our execution of the Perceptron algorithm is $M_{H_w} = \frac{1}{\sqrt{5}}$, which completes part (c).

(d) From lecture, we know that, for any training data $(X_1, Y_1), ..., (X_n, Y_n) \in \mathbb{R}^d \times \{\pm 1\}$ and any vector w^* such that $||w^*|| = 1$ and $(Y_j < w^*, X_j >) > 0$ for all $j \in \{1, ..., n\}$, the Perceptron algorithm is guaranteed to terminate in at most

$$T \leq \frac{R^2}{\gamma^2} \qquad (34)$$

iterations, where

$$R := \max_{j \in \{1,\ldots,n\}} ||X_j|| \qquad (35)$$

and

$$\gamma := \min_{j \in \{1,\ldots,n\}} Y_j < w^*, X_j > \qquad (36)$$

We can directly compute that

$$\begin{aligned} ||X_1|| &= \sqrt{1^2 + 0^2} = \sqrt{1} = 1 \\ ||X_2|| &= \sqrt{2^2 + 0^2} = \sqrt{4} = 2 \\ ||X_3|| &= \sqrt{0^2 + 1^2} = \sqrt{1} = 1 \\ ||X_4|| &= \sqrt{1^2 + 1^2} = \sqrt{2} \end{aligned}$$

so, for the given training data, we have

$$R = \max_{j \in \{1,2,3,4\}} ||X_j|| = ||X_2|| = 2 \qquad (37)$$

Note that since (34) provides an upper bound on the *maximum* number of iterations for the Perceptron algorithm to converge, we can let $T :=$ the *maximum* number of iterations needed for the Perceptron algorithm to converge given the training data from (24). Then, since our execution of the Perceptron algorithm with the training data from (24) took 5 iterations to converge, we know

$$T \geq 5 \qquad (38)$$

Combining (38) and (34) yields

$$5 \leq T \leq \frac{R^2}{\gamma^2} \qquad (39)$$

Rearranging (39) to get an upper bound for γ yields

$$\gamma^2 \leq \frac{R^2}{5} \implies \gamma \leq \frac{R}{\sqrt{5}} \qquad (40)$$

Plugging the result from (37) into (40) yields

$$\gamma \leq \frac{2}{\sqrt{5}} \qquad (41)$$

Thus, for the given training data from (24), all vectors w^* of magnitude $||w^*|| = 1$ that satisfy $(Y_j < w^*, X_j >) > 0$ for all $j \in \{1, 2, 3, 4\}$ also satisfy

$$\gamma := \min_{j \in \{1,2,3,4\}} Y_j < w^*, X_j > \leq \frac{2}{\sqrt{5}} \qquad (42)$$

Since $Y_j \in \{\pm 1\}$ and $(Y_j < w^*, X_j >) > 0$ for all $j \in \{1, 2, 3, 4\}$, we know

$$(Y_j < w^*, X_j >) = | < w^*, X_j > | \qquad (43)$$

for all $j \in \{1, 2, 3, 4\}$. Combining (43) and (42) yields

$$\min_{j \in \{1,2,3,4\}} | < w^*, X_j > | \leq \frac{2}{\sqrt{5}} \qquad (44)$$

From (26), we know that since $||w^*|| = 1$, the margin $M_{H_{w^*}}$ of the corresponding hyperplane $H_{w^*} = \{x :< w^*, x >= 0\}$ is

$$M_{H_{w^*}} = \min_{j\in\{1,2,3,4\}}(\text{ distance from } X_j \text{ to } H_{w^*}) = \min_{j\in\{1,2,3,4\}} | < w^*, X_j > | \qquad (45)$$

Combining (44) and (45), we find that for any hyperplane $H_{w^*} := \{x :< w^*, x >= 0\}$ corresponding to a vector w^* of magnitude $||w^*|| = 1$ that satisfies $Y_j(< w^*, X_j >) > 0$ for all $j \in \{1, ..., n\}$, the margin $M_{H_{w^*}}$ of H_{w^*} satisfies

$$M_{H_{w^*}} = \min_{j\in\{1,2,3,4\}}(\text{ distance from } X_j \text{ to } H_{w^*}) \leq \frac{2}{\sqrt{5}} \qquad (46)$$

From part (d), we find that

$$M_H \leq \frac{2}{\sqrt{5}}$$

is an upper bound for the maximum achievable margin M_H, which completes part (d).

3.

Exercise 5 in section 9.6 of the textbook.
(hint: does multiplying w by a positive constant t change the sign of the dot-proudct $< w, x >$?)
Suppose we modify the Perceptron algorithm as follows: In the update step, instead of performing $w^{(t+1)} = w^{(t)} + Y_i X_i$ whenever we make a mistake, we perform $w^{(t+1)} = w^{(t)} + \eta Y_i X_i$ for some $\eta > 0$. Prove that the modified Perceptron will perform the same number of iterations as the vanilla Perceptron and will converge to a vector that points to the same direction as the output of the vanilla Perceptron.

Solution. We want to show that the modified Perceptron algorithm both performs the same number of iterations as the vanilla Perceptron algorithm and converges to a vector that points to the same direction as the output of the vanilla algorithm.
Fix $\eta > 0$. Let $(X_1, Y_1), ..., (X_n, Y_n) \in \mathbb{R}^d \times \{\pm 1\}$ be the training data, and consider any arbitrary $w^* \in \mathbb{R}^d$ such that $||w^*|| = 1$ and $(Y_j < w^*, X_j >) > 0$ for all $j \in \{1, ..., n\}$. Define

$$\gamma := \min_{j\in\{1,...,n\}} Y_j < w^*, X_j >, \quad R := \max_{j\in\{1,...,n\}} ||X_j|| \qquad (47)$$

Then from lecture (and the hint from exercise 2 part (d)), we know the maximum number of iterations T for the vanilla Perceptron algorithm satisfies

$$T \leq \frac{R^2}{\gamma^2} \qquad (48)$$

Let T_m be the maximum number of iterations the modified Perceptron algorithm takes to converge for the given training data. Note that the specific number of iterations both the modified and vanilla Perceptron algorithms take to converge for the given training data depend on the order in which the training data are read by the algorithm when checking if $(Y_j < w^{(t)}, X_j >) \leq 0$ for all $j \in \{1, ..., n\}$ on the t'th iteration. Thus, we can also define N and N_m to be the number of iterations the vanilla and modified Perceptron algorithms, respectively, take to converge on the given training data permuted arbitrarily. Define $w^{(t)}$ to be the vanilla Perceptron's predicted vector at iteration t and $w_m^{(t)}$ to be the modified Perceptron's predicted vector at iteration t. Then, to complete the exercise, it suffices to show

$$(i) \quad w_m^{(t)} = \eta w^{(t)} \qquad (ii) \quad T_m \leq \frac{R^2}{\gamma^2} \qquad (iii) \quad N_m = N \qquad (49)$$

for all $1 \leq t \leq T, T_m$, any sufficient w^*, and any permutation of the training data.

First, we will show $w_m^{(t)} = \eta w^{(t)}$ for all $1 \leq t \leq T, T_m$. We will do so by strong induction on t.
Base Case: $t = 1$, then $w_m^{(t)} = (0,0) = w^{(t)}$, and $\eta(0,0) = (0,0)$, so $w_m^{(t)} = \eta w^{(t)}$ when $t = 1$.
Inductive Hypothesis: Assume that $w_m^{(t)} = \eta w^{(t)}$ for all $1 \leq t \leq k < T, T_m$.
Inductive Step: Consider $t = k + 1$.

Note: Since we go on to a $t = k + 1$'th iteration, we know there is some $j \in \{1, ..., n\}$ such that $(Y_j < w_m^{(k)}, X_j >) \leq 0$. Since $\eta > 0$, and $w_m^{(k)} = \eta w^{(k)}$ by the *Inductive Hypothesis*, we have

$$Y_j < w_m^{(k)}, X_j >= Y_j < \eta w^{(k)}, X_j >= \eta(Y_j < w^{(k)}, X_j >) \leq 0 \iff (Y_j < w^{(k)}, X_j >) \leq 0 \qquad (50)$$

The result from (50) will be instrumental for the remainder of the proof.

By the recursive definition of $w_m^{(k+1)}$ for the modified Perceptron algorithm, we have

$$w_m^{(k+1)} = w_m^{(k)} + \eta Y_j X_j \qquad (51)$$

Applying the inductive hypothesis to (51) yields

$$w_m^{(k+1)} = \eta w^{(k)} + \eta Y_j X_j = \eta(w^{(k)} + Y_j X_j) \qquad (52)$$

From (50), since $(Y_j < w_m^{(k)}, X_j >) \leq 0$, we know

$$(Y_j < w^{(k)}, X_j >) \leq 0 \qquad (53)$$

so by (52) and the recursive definition of $w^{(k+1)}$ for the vanilla Perceptron algorithm, assuming the training data are read in the same order by both algorithms, we have

$$w_m^{(k+1)} = \eta(w^{(k+1)}) \qquad (54)$$

The conclusion that $w_m^{(t)} = \eta w^{(t)}$ for all $1 \leq t \leq T, T_m$ follows by induction.

Now, we will show that $T_m \leq \frac{R^2}{\gamma^2}$. By definition, for all $\theta \in \mathbb{R}$, we have

$$cos(\theta) \leq 1 \qquad (55)$$

Using the result from (55) in combination with the definition of the dot product, we find that for any two vectors u and v (of nonzero magnitude)

$$< u, v >= ||u|| \cdot ||v|| \cdot cos(\theta) \implies \frac{< u, v >}{||u|| \cdot ||v||} = cos(\theta) \leq 1 \qquad (56)$$

where θ is the angle between u and v.
Define $w^{`} = \frac{w^*}{\gamma}$, so we have

$$(Y_j < w^{`}, X_j >) \geq 1 \quad \text{for all } j \in \{1, ..., n\} \qquad (57)$$

by the definition of γ. Then right hand side of (56) implies

$$1 \geq \frac{< w^{`}, w_m^{(t+1)} >}{||w^{`}|| \cdot ||w_m^{(t+1)}||} \qquad (58)$$

Claim: $(< w^{`}, w_m^{(t+1)} >) \geq \eta t$ for all $0 \leq t \leq T_m$.
Proof. We apply strong induction on t.
Base Case: $t = 0$, and $w_m^{(t+1)} = (0,0)$, so

$$(< w^{`}, w_m^{(t+1)}) = 0 \geq 0 = \eta t$$

so the claim holds for the base case.
Inductive Hypothesis: Assume $(< w^{`}, w_m^{(t+1)} >) \geq \eta t$ for all $0 \leq t \leq k-1 < T_m$.
Inductive Step: Consider $t = k$. Note that

$$(< w^{`}, w_m^{(k+1)} >) = (< w^{`}, w_m^{(k+1)} >) - (< w^{`}, w_m^{(k)} >) + (< w^{`}, w_m^{(k)} >) \tag{59}$$

and

$$(< w^{`}, w_m^{(k+1)} >) - (< w^{`}, w_m^{(k)} >) = (< w^{`}, w_m^{(k+1)} - w_m^{(k)} >) = (< w^{`}, \eta Y_j X_j >) = \eta Y_j (< w^{`}, X_j >) \tag{60}$$

From (57), we know that $(Y_j < w^{`}, X_j >) \geq 1$ for all $j \in \{1, ..., n\}$, so we have

$$(< w^{`}, w_m^{(k+1)} >) - (< w^{`}, w_m^{(k)} >) \geq \eta \tag{61}$$

Combining (59) with (61) yields

$$(< w^{`}, w_m^{(k+1)} >) \geq \eta + (< w^{`}, w_m^{(k)} >) \tag{62}$$

Applying the *Inductive Hypothesis* to (62) yields

$$(< w^{`}, w_m^{(k+1)} >) \geq \eta + \eta(k-1) = \eta k \tag{63}$$

The conclusion that $(< w^{`}, w_m^{(t+1)} >) \geq \eta t$ for all $0 \leq t \leq T_m$ follows from (63) by induction.

Claim: $||w_m^{(t+1)}||^2 \leq t\eta^2 R^2$ for all $0 \leq t \leq T_m$.
Proof: We apply strong induction on t.
Base Case: $t = 0$, we have $w_m^{(t+1)} = w_m^{(1)} = (0, 0)$, so

$$||w_m^{(t+1)}||^2 = 0 \leq 0 = 0\eta^2 R^2$$

so the claim holds for the base case.
Inductive Hypothesis: Assume $||w_m^{(t+1)}||^2 \leq t\eta^2 R^2$ for all $0 \leq t \leq k-1 < T_m$.
Inductive Step: Consider $t = k$. Then $w_m^{(t+1)} = w_m^{(k+1)}$, and

$$||w_m^{(k+1)}||^2 = ||w_m^{(k)} + \eta Y_j X_j||^2 = ||w_m^{(k)}||^2 + \eta^2 Y_j^2 ||X_j||^2 + 2\eta Y_j (< w_m^{(k)}, X_j >) \tag{64}$$

By assumption, we have $Y_j (< w_m^{(k)}, X_j >) \leq 0 \implies 2\eta Y_j (< w_m^{(k)}, X_j >) \leq 0$, since $\eta > 0$. Also, we have $Y_j^2 = 1$ since $Y_j \in \{\pm 1\}$ and $||X_j||^2 \leq R^2$ by the definition of R. Thus, we can rewrite (64) as

$$||w_m^{(k+1)}||^2 \leq ||w_m^{(k)}||^2 + \eta^2 R^2 \tag{65}$$

Applying the inductive hypothesis to (65) yields

$$||w_m^{(k+1)}||^2 \leq (k-1)\eta^2 R^2 + \eta^2 R^2 = k\eta^2 R^2 \tag{66}$$

The conclusion that $||w_m^{(t+1)}||^2 \leq t\eta^2 R^2$ for all $0 \leq t \leq T_m$ follows from (66) by induction.

Also, since $w^{`} := \frac{w^*}{\gamma}$, we know

$$||w^{`}|| = \frac{1}{\gamma} ||w^*|| = \frac{1}{\gamma} \tag{67}$$

with the last equality following since $||w^*|| = 1$ by definition.

Plugging the results from (63), (66), and (67) into (58), we see

$$1 \geq \frac{< w^{`}, w_m^{(t+1)} >}{||w^{`}|| \cdot ||w_m^{(t+1)}||} \geq \frac{\eta t}{\frac{\sqrt{t\eta^2 R^2}}{\gamma}} = \frac{\sqrt{t}\gamma}{R} \tag{68}$$

Since (68) holds for all $0 \leq t \leq T_m$, we know

$$1 \geq \frac{\sqrt{T_m}\gamma}{R} \implies \frac{R}{\gamma} \geq \sqrt{T_m} \implies T_m \leq \frac{R^2}{\gamma^2} \qquad (69)$$

The result from (69) implies that the number of iterations it takes the modified Perceptron algorithm to converge is always upper bounded by the exact same upper bound as for the vanilla Perceptron algorithm.

Now, we will show that $N_m = N$ for any given permutation of the training data. Assume to the contrary that $N_m \neq N$ for some permutation of the training data. Then either $N_m < N$ or $N_m > n$. If $N_m < N$, then at iteration $t = N_m$, we must have $Y_j(< w_m^{(N_m)}, X_j >) > 0$ for all $j \in \{1, ..., n\}$ but there must be some $i \in \{1, ..., n\}$ such that $Y_i(< w^{(N_m)}, X_i >) \leq 0$ so the vanilla Perceptron algorithm can update $w^{(N_m+1)}$ for the next iteration. However, from (50), we know that for any permutation of the training data, we have

$$Y_j(< w_m^{(N_m)}, X_j >) \leq 0 \iff Y_j(< w^{(N_m)}, X_j >) \leq 0$$

so we have a contradiction.
Similarly, if $N_m > N$, then at iteration $t = N$, we must have $Y_j(< w^{(N)}, X_j >) > 0$ for all $j \in \{1, ...n\}$ but there must be some $i \in \{1, ..., n\}$ such that $Y_j(< w_m^{(N)}, X_j >) \leq 0$ so that the modified Perceptron algorithm can update $w^{(N+1)}$ for the next iteration. Once again, this produces a contradiction with (50).
Thus, if $N_m \neq N$, we can always derive a contradiction, so we know $N_m = N$. That is, for any given ordering of the training data, the modified Perceptron algorithm will *always* converge in the *same* number of iterations as the vanilla Perceptron algorithm. This combines with (54) to imply that

$$w_m^{(N_m)} = w_m^{(N)} = \eta w^{(N_m)} = \eta w^{(N)} \qquad (70)$$

for any possible number of iterations $N = N_m$ required by both the vanilla and modified Perceptron algorithms to converge. Since the modified Perceptron algorithm will return $w_m^{(N_m)}$ by the definition of N_m and the vanilla Perceptron algorithm will return $w^{(N)}$ by the definition of N, (70) combines with the fact that $\eta > 0$ to imply that the modified and vanilla Perceptron algorithms will *always* converge to vectors that point in the same direction, for any given permutation of training data.
This completes the proof that the modified Perceptron algorithm will perform the same number of iterations as the vanilla Perceptron and will converge to a vector that points in the same direction as the output of the vanilla Perceptron.

Assignment 5

Read chapter 10 of the textbook "Understanding Machine Learning." Then do the following problems:

1. Logistic regression and the Bayes classifier.

Recall that the Logistic Regression solves the following problem:

$$\frac{1}{n}\sum_{j=1}^{n}\big(ln(1 + e^{\langle w, X_j\rangle}) - Y_j\langle w, X_j\rangle\big) \rightarrow \text{ minimize over } w \in \mathbb{R}^d$$

where $(X_1, Y_1), ..., (X_n, Y_n) \in \mathbb{R}^d \times \{0, 1\}$ is the training data. Assume that we know the distribution of (X, Y) so that we can replace the average by the expectation:

$$\mathbb{E}[ln(1 + e^{\langle w, X\rangle}) - Y\langle w, X\rangle] \rightarrow \text{ minimize over } w \in \mathbb{R}^d$$

Let us replace the linear function $\langle w, X\rangle$ by an arbitrary, possibly non-linear function $g(X)$, and let g_* be the function that minimizes

$$\mathbb{E}[ln(1 + e^{g(X)}) - Yg(X)]$$

over all functions g(X). In the course of this exercise, we will show that $g_*(X)$ gives rise to the Bayes classifier. In class, we followed the same pattern to justify the AdaBoost algorithm.

(a) If the labels take values $\{0, 1\}$, instead of $\{-1, +1\}$, show that the expression for the Bayes classifier g_* in terms of the conditional probability $p(x) := \mathbb{P}(Y = 1|X = x)$ is

$$g_*(x) = \begin{cases} 1 \text{ if } p(x) \geq \frac{1}{2} \\ 0 \text{ if } p(x) < \frac{1}{2} \end{cases}$$

(b) Recall that in the logistic regression framework discussed above, the estimator of the probability $p(x)$ is constructed from the vector w as $\frac{e^{\langle w,x\rangle}}{1+e^{\langle w,x\rangle}}$. Show that the optimal function $g_*(x)$ (meaning the one that minimizes $\mathbb{E}[ln(1 + e^{g(X)}) - Yg(X)]$ is such that

$$\frac{e^{g_*(x)}}{1 + e^{g_*(x)}} = p(x)$$

and conclude that the associated binary classifier is the Bayes classifier.
You can assume that X takes values $x_1, ..., x_k$ with probabilities $p_1, ..., p_k$, respectively, and follow the steps of the proof of a similar result for AdaBoost from the class notes.

(c) How would you interpret the result established above?

Solution.

(a) By definition, the Bayes classifier is the function $g : \mathbb{R}^d \to \{0, 1\}$ which minimizes $\mathbb{P}(Y \neq g(X))$ over all possible training data. Note that, for any $x \in \mathbb{R}^d$, either $p(x) > \frac{1}{2}$, $p(x) < \frac{1}{2}$, or $p(x) = \frac{1}{2}$. We will show that, in each case, the classifier g_*, defined as

$$g_*(x) := \begin{cases} 1 \text{ if } p(x) \geq \frac{1}{2} \\ 0 \text{ if } p(x) < \frac{1}{2} \end{cases} \qquad (1)$$

returns the value $g_*(x) \in \{0, 1\}$ that minimizes $\mathbb{P}(Y \neq g_*(X)|X = x)$. We will do so by showing, in each case, that the output of g_* as defined in (1) produces a lower probability of classification error than the alternative output.
First, consider any $x \in \mathbb{R}^d$ such that $p(x) > \frac{1}{2}$. Then $g_*(x) = 1$ by definition. Thus,

$$\mathbb{P}(Y \neq g_*(X)|X = x) = \mathbb{P}(Y = 0|X = x) = 1 - \mathbb{P}(Y = 1|X = x) = 1 - p(x) < 1 - \frac{1}{2} = \frac{1}{2} \qquad (2)$$

Since g_* maps to $\{0, 1\}$, the only other possible outcome when $p(x) > \frac{1}{2}$ is $g_*(x) = 0$. However, if $g_*(x) = 0$, we can easily compute that

$$\mathbb{P}(Y \neq 0|X = x) = \mathbb{P}(Y = 1|X = x) =: p(x) > \frac{1}{2} \qquad (3)$$

Comparing (2) and (3), we see that $g_*(x) = 1$ minimizes $\mathbb{P}(Y \neq g_*(X)|X = x)$ for all $x \in \mathbb{R}^d$ such that $p(x) > \frac{1}{2}$.
Next, consider any $x \in \mathbb{R}^d$ such that $p(x) < \frac{1}{2}$. From (1), we see that $g_*(x) = 0$ by definition. This yields

$$\mathbb{P}(Y \neq g_*(X)|X = x) = \mathbb{P}(Y = 1|X = x) = p(x) < \frac{1}{2} \qquad (4)$$

If g_* does not output 0 for any such x, it must output 1, so we would have

$$\mathbb{P}(Y \neq 1|X = x) = 1 - \mathbb{P}(Y = 1|X = x) = 1 - p(x) > \frac{1}{2} \qquad (5)$$

Comparing (4) and (5), we see that $g_*(x) = 0$ minimizes $\mathbb{P}(Y \neq g_*(X)|X = x)$ for all $x \in \mathbb{R}^d$ such that $p(x) < \frac{1}{2}$. Finally consider all $x \in \mathbb{R}^d$ such that $p(x) = \frac{1}{2}$. From (1), we know that $g_*(x) = 1$ by definition. This yields

$$\mathbb{P}(Y \neq g_*(X)|X = x) = \mathbb{P}(Y = 0|X = x) = 1 - \mathbb{P}(Y = 1|X = x) = 1 - \frac{1}{2} = \frac{1}{2} \qquad (6)$$

If, on the other hand, $g_*(x) = 0$ for such x, we would have

$$\mathbb{P}(Y \neq 0|X = x) = \mathbb{P}(Y = 1|X = x) = p(x) = \frac{1}{2} \qquad (7)$$

Comparing (6) and (7), we see that, for all $x \in \mathbb{R}^d$ such that $p(x) = \frac{1}{2}$, there is no difference between the probability of classification error for classifiers outputting 1 and those outputting 0. Thus, either output, including the $g_*(x) = 1$ defined by (1), minimize $\mathbb{P}(Y \neq g_*(X)|X = x)$ for all such x.
In every case, g_* as defined in (1) produces an output with a lower generalization error than the alternative output. Therefore, g_* minimizes $\mathbb{P}(Y \neq g(X)|X = x)$ over all $x \in \mathbb{R}^d$ and all binary classifiers g, so g_* minimizes $\mathbb{P}(Y \neq g(X))$ over the distribution of the training data and all possible binary classifiers g. By definition, this means g_* is the Bayes classifier for the label set $\{0, 1\}$.

(b) We follow the steps of the proof of a similar result for AdaBoost from lecture. We are given that X is discrete, taking values in $\{x_1, ..., x_k\}$ with corresponding probabilities in $\{p_1, ..., p_k\}$. By the law of total expectation, since

$$(X = x_1), \cdots, (X = x_k)$$

are all mututally disjoint events whose union is the sample space, we have

$$\mathbb{E}[ln(1 + e^{g(X)}) - Yg(X)] = \sum_{i=1}^{k} \mathbb{E}[\big(ln(1 + e^{g(X)}) - Yg(X)\big)|X = x_i] \cdot p_i \qquad (8)$$

For each term inside the sum, if we are given $X = x_i$, then we know $g(X) = g(x_i)$, so $g(X)$ is a constant. Thus, given $X = x_i$, $ln(1 + e^{g(X)})$ is also a constant, so we can pull it out of the expectation to find

$$\mathbb{E}[ln(1 + e^{g(X)}) - Yg(X)] = \sum_{i=1}^{k} (ln(1 + e^{g(x_i)}) - \mathbb{E}[Yg(X)|X = x_i]) \cdot p_i \qquad (9)$$

Since $g(X) = g(x_i)$ is constant given $X = x_i$, and $\mathbb{E}[aY|Z] = a\mathbb{E}[Y]$ for all constants $a \in \mathbb{R}$, we can also pull the $g(X)$ out of the expectation to find

$$\mathbb{E}[ln(1 + e^{g(X)}) - Yg(X)] = \sum_{i=1}^{k} (ln(1 + e^{g(x_i)}) - g(x_i)\mathbb{E}[Y|X = x_i]) \cdot p_i \qquad (10)$$

We know $Y \in \{0, 1\}$, so we can easily compute that

$$\mathbb{E}[Y|X = x_i] = \sum_{y=0}^{1} y\mathbb{P}(Y = y|X = x_i) = 0\mathbb{P}(Y = 0|X = x_i) + 1\mathbb{P}(Y = 1|X = x_i) = p(x_i) \qquad (11)$$

The result from (11) allows us to remove the expectation entirely from (10):

$$\mathbb{E}[ln(1 + e^{g(X)}) - Yg(X)] = \sum_{i=1}^{k} (ln(1 + e^{g(x_i)}) - g(x_i)p(x_i)) \cdot p_i \qquad (11)$$

Now, define $t_1, ..., t_k$ such that $t_i = g(x_i)$ for all $i \in \{1, ..., k\}$. From (11), we see that to minimize $\mathbb{E}[ln(1 + e^{g(X)}) - Yg(X)]$, it suffices to minimize

$$f(t_i) := ln(1 + e^{t_i}) - t_i p(x_i) \qquad (12)$$

Differentiating with respect to t_i yields

$$f'(t_i) = \frac{d}{dt}f(t_i) = \frac{e^{t_i}}{1+e^{t_i}} - p(x_i) \qquad (13)$$

To find a critical point on f, we just need to find a t_i that satisfies $f'(t_i) = 0$. We can easily compute that

$$f'(t_i) = \frac{e^{t_i}}{1+e^{t_i}} - p(x_i) = 0 \iff \frac{e^{t_i}}{1+e^{t_i}} = p(x_i) \qquad (14)$$

Also, by differentiating (13) once more with respect to t_i, we find

$$f"(t_i) = \frac{d}{dt}f'(t_i) = \frac{(1+e^{t_i})e^{t_i} - e^{t_i}e^{t_i}}{(1+e^{t_i})^2} = \frac{e^{t_i}}{(1+e^{t_i})^2} \qquad (15)$$

Since e^x is a non-negative function, we have

$$f"(t_i) > 0 \qquad (16)$$

so we know the critical point t_i, which must satisfy (14), is indeed a global minimum of the function f. Substituting $g(x_i)$ for t_i in (14), we find that the optimal g_* which minimizes $\mathbb{E}[ln(1+e^{g(X)}) - Yg(X)]$ satisfies

$$\frac{e^{g_*(x_i)}}{1+e^{g_*(x_i)}} = p(x_i) \qquad (17)$$

for all $i \in \{1, ..., k\}$, so we have

$$\frac{e^{g_*(x)}}{1+e^{g_*(x)}} = p(x) \qquad (18)$$

The binary classifier associated with this function g_* is defined to be

$$g_*^b(x) := \begin{cases} 1 \text{ if } \frac{e^{g_*(x)}}{1+e^{g_*(x)}} \geq \frac{1}{2} \\ 0 \text{ otherwise.} \end{cases} \qquad (19)$$

Using the result from (18), we can simplify (19) to be

$$g_*^b(x) = \begin{cases} 1 \text{ if } p(x) \geq \frac{1}{2} \\ 0 \text{ if } p(x) < \frac{1}{2} \end{cases} \qquad (20)$$

Comparing (20) and (1), we see that this binary classifier g_*^b corresponding to the function g_* that minimizes $\mathbb{E}[ln(1+e^{g(X)}) - Yg(X)]$ is actually just the Bayes classifier itself. That is

$$\underbrace{g_*^b(x)}_{\text{the binary classifier corresponding to the optimal function } g_* \text{ defined in (17)}} = \underbrace{g_*(x)}_{\text{the binary classifier (Bayes classifier) defined in (1)}} \qquad (21)$$

for all x.

(c) I interpret the results established above to mean that Logistic Regression can identify and output a function g_* which is sufficient to compute the Bayes classifier directly, without prior explicit knowledge about the conditional distribution of Y given X. Thus, Logistic Regression appears to be a powerful tool for extracting ideal binary classifiers from problems with unknown distributions over the training data. By definition, the Bayes classifier performs the best over the true distribution of the training data. Thus, obtaining it without needing to explicitly define that distribution could allow Logistic Regression to solve problems over complicated distributions which cannot be explicitly defined. Moreover, since the solution from Logistic Regression can yield the Bayes classifier, the algorithm not only can solve difficult binary classification problems, but it can solve them ideally (or at least approximately ideally).

2. AdaBoost.

Recall that on ever step t, the AdaBoost algorithm generates weights $w_1^{(t+1)}, ..., w_n^{(t+1)}$ attached to the samples $(X_1, Y_1), ..., (X_n, Y_n)$. The goal of this problem is to understand how exactly these weights change. Recall from the class notes that

$$w_j^{(t+1)} = \frac{w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j)}{Z_t}$$

where

$$Z_t := \sum_{i=1}^{n} w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j)), \quad \alpha_t := \frac{1}{2} ln(\frac{1 - e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)}, \quad \text{and } e_{n,w^{(t)}}(f_t) := \sum_{j=1}^{n} w_j^{(t)} I\{Y_j \neq f_t(X_j)\}$$

(a) Read the class notes and prove that

$$Z_t = 2\sqrt{e_{n,w^{(t)}}(f_t)(1 - e_{n,w^{(t)}}(f_t))}$$

This was already done in class, but you need to fill in the missing details.

(b) Simplify the expression for $w_j^{(t+1)}$ using this expression for Z_t (consider two cases seprately: $Y_j = f_t(X_j)$ and $Y_j \neq f_t(X_j)$).

(c) Recall that, by the "weak learnability" assumption of AdaBoost, $e_{n,w^{(t)}}(f_t) < \frac{1}{2}$. Use this fact together with the simplified expression for $w_j^{(t+1)}$ to show that, if f_t classifies X_j correctly, then $w_j^{(t+1)} < w_j^{(t)}$, otherwise $w_j^{(t+1)} > w_j^{(t)}$.

Solution.

(a) Note that, for all $j \in \{1, ..., n\}$,

$$I\{Y_j = f_t(X_j)\} = 1 \iff I\{Y_j \neq f_t(X_j)\} = 0 \quad \text{and} \quad I\{Y_j = f_t(X_j)\} = 0 \iff I\{Y_j \neq f_t(X_j)\} = 1 \tag{22}$$

which directly implies

$$I\{Y_j = f_t(X_j)\} + I\{Y_j \neq f_t(X_j)\} = 1 \tag{23}$$

for all $j \in \{1, ..., n\}$. Thus, we can write

$$Z_t := \sum_{j=1}^{n} w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j)) = \sum_{j=1}^{n} w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j))(I\{Y_j = f_t(X_j)\} + I\{Y_j \neq f_t(X_j)\}) \tag{24}$$

Adding and subtracting $\sum_{j=1}^{n} w_j^{(t)} e^{-\alpha t} I\{Y_j \neq f_t(X_j)\}$ to (24) yields

$$\begin{aligned} Z_t = & \sum_{j=1}^{n} w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j)) I\{Y_j = f_t(X_j)\} + \sum_{j=1}^{n} w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j)) I\{Y_j \neq f_t(X_j)\}) \\ + & \sum_{j=1}^{n} w_j^{(t)} e^{-\alpha t} I\{Y_j \neq f_t(X_j)\} - \sum_{j=1}^{n} w_j^{(t)} e^{-\alpha t} I\{Y_j \neq f_t(X_j)\} \end{aligned} \tag{25}$$

If $Y_j \neq f_t(X_j)$, then

$$w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j)) I\{Y_j = f_t(X_j)\} = 0 \tag{26}$$

and if $Y_j = f_t(X_j)$, then $Y_j f_t(X_j) = 1$, so

$$w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j)) I\{Y_j = f_t(X_j)\} = w_j^{(t)} e^{-\alpha_t} \tag{27}$$

Whenever $Y_j = f_t(X_j)$, (27) holds, so we can write

$$\sum_{j=1}^{n} w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j)) I\{Y_j = f_t(X_j)\} = \sum_{j=1}^{n} w_j^{(t)} e^{-\alpha_t} I\{Y_j = f_t(X_j)\} \tag{28}$$

Similarly, if $Y_j = f_t(X_j)$, then

$$w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j)) I\{Y_j \neq f_t(X_j)\} = 0 \tag{29}$$

while $Y_j \neq f_t(X_j)$ implies $Y_j f_t(X_j) = -1$ and, subsequently,

$$exp(-\alpha_t Y_j f_t(X_j)) I\{Y_j \neq f_t(X_j)\} = w_j^{(t)} e^{\alpha_t} \tag{30}$$

This allows us to write

$$\sum_{j=1}^{n} w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j)) I\{Y_j \neq f_t(X_j)\} = \sum_{j=1}^{n} w_j^{(t)} e^{\alpha_t} I\{Y_j \neq f_t(X_j)\} \tag{31}$$

Plugging the results from (28) and (31) into (25) yields

$$\begin{aligned} Z_t = & \sum_{i=1}^{n} w_j^{(t)} e^{-\alpha_t} I\{Y_j = f_t(X_j)\} + \sum_{j=1}^{n} w_j^{(t)} e^{\alpha_t} I\{Y_j \neq f_t(X_j)\} \\ + & \sum_{j=1}^{n} w_j^{(t)} e^{-\alpha t} I\{Y_j \neq f_t(X_j)\} - \sum_{j=1}^{n} w_j^{(t)} e^{-\alpha t} I\{Y_j \neq f_t(X_j)\} \end{aligned} \tag{32}$$

Rearranging the terms in (32), and pulling the e^{α_t} and $e^{-\alpha_t}$ terms out of the summations, we find

$$Z_t = (e^{\alpha_t} - e^{-\alpha_t}) \sum_{j=1}^{n} w_j^{(t)} I\{Y_j \neq f_t(X_j)\} + e^{-\alpha_t} \sum_{j=1}^{n} w_j^{(t)} (I\{Y_j = f_t(X_j)\} + I\{Y_j \neq f_t(X_j)\}) \tag{33}$$

Plugging the result from (23) into (33) yields

$$Z_t = (e^{\alpha_t} - e^{-\alpha_t}) \sum_{j=1}^{n} w_j^{(t)} I\{Y_j \neq f_t(X_j)\} + e^{-\alpha_t} \sum_{i=1}^{n} w_j^{(t)} \tag{34}$$

By definition, for any t, $\sum_{j=1}^{n} w_j^{(t)} = 1$, so we have

$$Z_t = e^{-\alpha_t} + (e^{\alpha_t} - e^{-\alpha_t}) \sum_{j=1}^{n} w_j^{(t)} I\{Y_j \neq f_t(X_j)\} \tag{35}$$

Noting that the summation from (35) is just the definition of $e_{n,w^{(t)}}(f_t)$ from the problem statement, we find

$$Z_t = e^{-\alpha_t} + (e^{\alpha_t} - e^{-\alpha_t}) e_{n,w^{(t)}}(f_t) \tag{36}$$

Now, we will use the definition of α_t to simplify the equation for Z_t further. By the definition from the problem statement, we have

$$e^{-\alpha_t} = e^{-\frac{1}{2} ln(\frac{1-e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)})} = \big(e^{ln(\frac{1-e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)})}\big)^{-\frac{1}{2}} = \big(\frac{1-e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)}\big)^{-\frac{1}{2}} = \sqrt{\frac{e_{n,w^{(t)}}(f_t)}{1-e_{n,w^{(t)}}(f_t)}} \tag{37}$$

and

$$e^{\alpha_t} = e^{\frac{1}{2} ln(\frac{1-e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)})} = \big(e^{ln(\frac{1-e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)})}\big)^{\frac{1}{2}} = \big(\frac{1-e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)}\big)^{\frac{1}{2}} = \sqrt{\frac{1-e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)}} \tag{38}$$

Plugging the results from (37) and (38) into (36) and simplifying yields

$$\begin{aligned}
Z_t = & \sqrt{\frac{e_{n,w^{(t)}}(f_t)}{1-e_{n,w^{(t)}}(f_t)}} + \left(\sqrt{\frac{1-e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)}} - \sqrt{\frac{e_{n,w^{(t)}}(f_t)}{1-e_{n,w^{(t)}}(f_t)}}\right) e_{n,w^{(t)}}(f_t) \\
= & \sqrt{\frac{e_{n,w^{(t)}}(f_t)}{1-e_{n,w^{(t)}}(f_t)}} + \sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t)} - \frac{e_{n,w^{(t)}}(f_t)\sqrt{e_{n,w^{(t)}}(f_t)}}{\sqrt{1-e_{n,w^{(t)}}(f_t)}} \\
= & \sqrt{\frac{e_{n,w^{(t)}}(f_t)}{1-e_{n,w^{(t)}}(f_t)}}\sqrt{\frac{e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)}} + \sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t)} \cdot \frac{\sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t)}}{\sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t)}} \\
- & \frac{e_{n,w^{(t)}}(f_t)\sqrt{e_{n,w^{(t)}}(f_t)}}{\sqrt{1-e_{n,w^{(t)}}(f_t)}}\sqrt{\frac{e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)}} \\
= & \frac{e_{n,w^{(t)}}(f_t) + e_{n,w^{(t)}}(f_t)\cdot(1-e_{n,w^{(t)}}(f_t)) - e_{n,w^{(t)}}(f_t)^2}{\sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t))}} \\
= & \frac{e_{n,w^{(t)}}(f_t)\cdot(1-e_{n,w^{(t)}}(f_t)) + e_{n,w^{(t)}}(f_t)\cdot(1-e_{n,w^{(t)}}(f_t))}{\sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t))}} \\
= & \frac{2e_{n,w^{(t)}}(f_t)\cdot(1-e_{n,w^{(t)}}(f_t))}{\sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t))}} \\
= & 2\sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t))} \qquad (39)
\end{aligned}$$

The last line of (39) completes the proof that

$$Z_t = 2\sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t))}$$

which completes part (a).

(b) To simplify the expression for $w_j^{(t+1)}$, we plug in the expression for Z_t from (39) and use (23) to find

$$\begin{aligned}
w_j^{(t+1)} = & \frac{w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j))}{2\sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t))}}(I\{Y_j = f_t(X_j)\} + I\{Y_j \neq f_t(X_j)\}) \\
= & \frac{w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j))}{2\sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t))}} I\{Y_j = f_t(X_j)\} + \frac{w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j))}{2\sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t))}} I\{Y_j \neq f_t(X_j)\} \qquad (40)
\end{aligned}$$

We already showed that

$$w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j)) I\{Y_j = f_t(X_j)\} = w_j^{(t)} e^{-\alpha_t} I\{Y_j = f_t(X_j)\} \qquad (41)$$

in (27) and

$$w_j^{(t)} exp(-\alpha_t Y_j f_t(X_j)) I\{Y_j \neq f_t(X_j)\} = w_j^{(t)} e^{\alpha_t} I\{Y_j \neq f_t(X_j)\} \qquad (42)$$

in (30). Plugging these results into (40) yields

$$w_j^{(t+1)} = \frac{w_j^{(t)} e^{-\alpha_t}}{2\sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t))}} I\{Y_j = f_t(X_j)\} + \frac{w_j^{(t)} e^{\alpha_t}}{2\sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t))}} I\{Y_j \neq f_t(X_j)\} \qquad (43)$$

Plugging the results from (37) and (38) into (43) allows us to further simplify the equation for $w_j^{(t+1)}$:

$$\begin{aligned} w_j^{(t+1)} &= \frac{w_j^{(t)}\sqrt{\frac{e_{n,w^{(t)}}(f_t)}{1-e_{n,w^{(t)}}(f_t)}}}{2\sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t))}} I\{Y_j = f_t(X_j)\} + \frac{w_j^{(t)}\sqrt{\frac{1-e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)}}}{2\sqrt{e_{n,w^{(t)}}(f_t)(1-e_{n,w^{(t)}}(f_t))}} I\{Y_j \neq f_t(X_j)\} \\ &= \frac{w_j^{(t)}}{2}\frac{1}{1-e_{n,w^{(t)}}(f_t)} I\{Y_j = f_t(X_j)\} + \frac{w_j^{(t)}}{2}\frac{1}{e_{n,w^{(t)}}(f_t)} I\{Y_j \neq f_t(X_j)\} \\ &= \frac{w_j^{(t)}}{2(1-e_{n,w^{(t)}}(f_t))} I\{Y_j = f_t(X_j)\} + \frac{w_j^{(t)}}{2e_{n,w^{(t)}}(f_t)} I\{Y_j \neq f_t(X_j)\} \\ &= \frac{w_j^{(t)}}{2}\left(\frac{I\{Y_j = f_t(X_j)\}}{1-e_{n,w^{(t)}}(f_t)} + \frac{I\{Y_j \neq f_t(X_j)\}}{e_{n,w^{(t)}}(f_t)}\right) \qquad (44) \end{aligned}$$

Thus, our simplified expression for $w_j^{(t+1)}$ is

$$w_j^{(t+1)} = \frac{w_j^{(t)}}{2}\left(\frac{I\{Y_j = f_t(X_j)\}}{1-e_{n,w^{(t)}}(f_t)} + \frac{I\{Y_j \neq f_t(X_j)\}}{e_{n,w^{(t)}}(f_t)}\right) \qquad (45)$$

By (22), if $Y_j = f_t(X_j)$, we have

$$w_j^{(t+1)} = \frac{w_j^{(t)}}{2}\frac{I\{Y_j = f_t(X_j)\}}{1-e_{n,w^{(t)}}(f_t)} = \frac{w_j^{(t)}}{2}\frac{1}{1-e_{n,w^{(t)}}(f_t)} \qquad (46)$$

and if $Y_j \neq f_t(X_j)$, we have

$$w_j^{(t+1)} = \frac{w_j^{(t)}}{2}\frac{I\{Y_j \neq f_t(X_j)\}}{e_{n,w^{(t)}}(f_t)} = \frac{w_j^{(t)}}{2}\frac{1}{e_{n,w^{(t)}}(f_t)} \qquad (47)$$

This completes part (b).

(c) It suffices to show:

(i) If $f_t(X_j) = Y_j$, then $w_j^{(t+1)} < w_j^{(t)}$.

(ii) If $f_t(X_j) \neq Y_j$, then $w_j^{(t+1)} > w_j^{(t)}$.

First we will show (i). From (46), if $Y_j = f_t(X_j)$, we have

$$w_j^{(t+1)} = \frac{w_j^{(t)}}{2(1-e_{n,w^{(t)}}(f_t))} \qquad (48)$$

Note that

$$e_{n,w^{(t)}}(f_t) < \frac{1}{2} \implies 1-e_{n,w^{(t)}}(f_t) > \frac{1}{2} \implies 2(1-e_{n,w^{(t)}}(f_t)) > 1 \implies \frac{1}{2(1-e_{n,w^{(t)}}(f_t))} < 1 \qquad (49)$$

Plugging the result from (49) into (48) yields

$$w_j^{(t+1)} = w_j^{(t)} \cdot \frac{1}{2(1-e_{n,w^{(t)}}(f_t))} < w_j^{(t)} \cdot 1 = w_j^{(t)} \qquad (50)$$

for all j such that $Y_j = f_t(X_j)$, which completes the proof of part (i).
Next, we show (ii). From (47), if $Y_j \neq f_t(X_j)$, we have

$$w_j^{(t+1)} = w_j^{(t)}\frac{1}{2e_{n,w^{(t)}}(f_t)} \qquad (51)$$

Note that

$$e_{n,w^{(t)}}(f_t) < \frac{1}{2} \implies 2e_{n,w^{(t)}}(f_t) < 1 \implies \frac{1}{2e_{n,w^{(t)}}(f_t)} > 1 \qquad (52)$$

Plugging the result from (52) into (51) yields

$$w_j^{(t+1)} = w_j^{(t)} \frac{1}{2e_{n,w^{(t)}}(f_t)} > w_j^{(t)} \cdot 1 = w_j^{(t)} \qquad (53)$$

for all j such that $Y_j \neq f_t(X_j)$, which completes the proof of part (ii). Thus, we have shown that

$$Y_j = f_t(X_j) \implies w_j^{(t+1)} < w_j^{(t)} \qquad (54)$$

and

$$Y_j \neq f_t(X_j) \implies w_j^{(t+1)} > w_j^{(t)} \qquad (55)$$

which completes part (c).

3. Expressive power of the convex combinations ("majority vote").

Recall that AdaBoost finds a solution in the class

$$G = \{\sum_{j=1}^{k} \alpha_j f_j : k \geq 1, \alpha_1, ..., \alpha_k \geq 0, f_1, ..., f_k \in \mathcal{F}\}$$

where $\mathcal{F}$ is some class of binary classifiers. How complex, or expressive, can the class of binary classifiers $\{sign(g), g \in G\}$ associated with G be? We look at one example in this exercise. Let $\mathcal{F}$ be the class of threshold classifiers, or "decision stumps,"

$$\mathcal{F} := \{f_{\theta,b}(x) = sign(x - \theta) \cdot b : \theta \in \mathbb{R}, b \in \{-1, +1\}\}$$

In other words, $f_{\theta,b}(x) = \begin{cases} b \text{ if } x \geq \theta, \\ -b \text{ if } x < \theta. \end{cases}$

(a) Let $-\infty = \theta_0 < \theta_1 < \cdots < \theta_r = \infty$ be a sequence of real numbers, and define

$$g_r(x) := \sum_{j=1}^{r} \alpha_j I\{x \in (\theta_{j-1}, \theta_j]\}$$

where $\alpha_j = (-1)^j$. Draw the graph of a generic function g_r for, say, $r = 4$ (you can pick θ's as you wish).

(b) Show that any function g_r can be realized as an element of the class G (after taking the sign). Specifically, let

$$h(x) = sign\big(\sum_{j=1}^{r} w_j \cdot sign(x - \theta_{j-1})\big)$$

where $w_1 = -\frac{1}{2}$ and $w_j = (-1)^j$ for all $j > 1$, and show that $h(x) = g_r(x)$.

Solution.

(a) We choose $r = 4$ and

$$-\infty = \theta_0 < -20 = \theta_1 < 0 = \theta_2 < 20 = \theta_3 < \theta_4 = \infty \qquad (56)$$

Note: By (56), for all $x \in \mathbb{R}$,

$$x \in (\theta_{j-1}, \theta_j] \implies x \notin (\theta_{i-1}, \theta_j] \text{ for all } i \neq j \qquad (57)$$

Note that, for all $x \in (-\infty, -20]$, we have

$$g_r(x) = \alpha_1 I\{x \in (-\infty, -20]\} + \sum_{j=2}^{4} \alpha_j I\{x \in (\theta_{j-1}, \theta_j)\} = \alpha_1 + 0 = \alpha_1 = -1 \qquad (58)$$

since $x \in (-\infty, -20]$ to implies that $I\{x \in (\theta_{j-1}, \theta)\} = 0$ for all $j > 1$ by (57).
For all $x \in (-20, 0]$, we have

$$g_r(x) = \alpha_2 I\{x \in (-20, 0]\} + \sum_{j \neq 2, j \in \{1,2,3,4\}} \alpha_j I\{x \in (\theta_{j-1}, \theta_j]\} = \alpha_2 I\{x \in (-20, 0]\} = \alpha_2 = (-1)^2 = 1 \qquad (59)$$

since $x \in (-20, 0]$ implies $x \notin (\theta_{j-1}, \theta_j]$ for all $j \neq 2$ by (57).
For all $x \in (0, 20]$, we have

$$g_r(x) = \alpha_3 I\{x \in (0, 20]\} + \sum_{j \neq 3, j \in \{1,2,3,4\}} \alpha_j I\{x \in (\theta_{j-1}, \theta_j)\} = \alpha_3 I\{x \in (0, 20]\} = \alpha_3 = (-1)^3 = -1 \qquad (60)$$

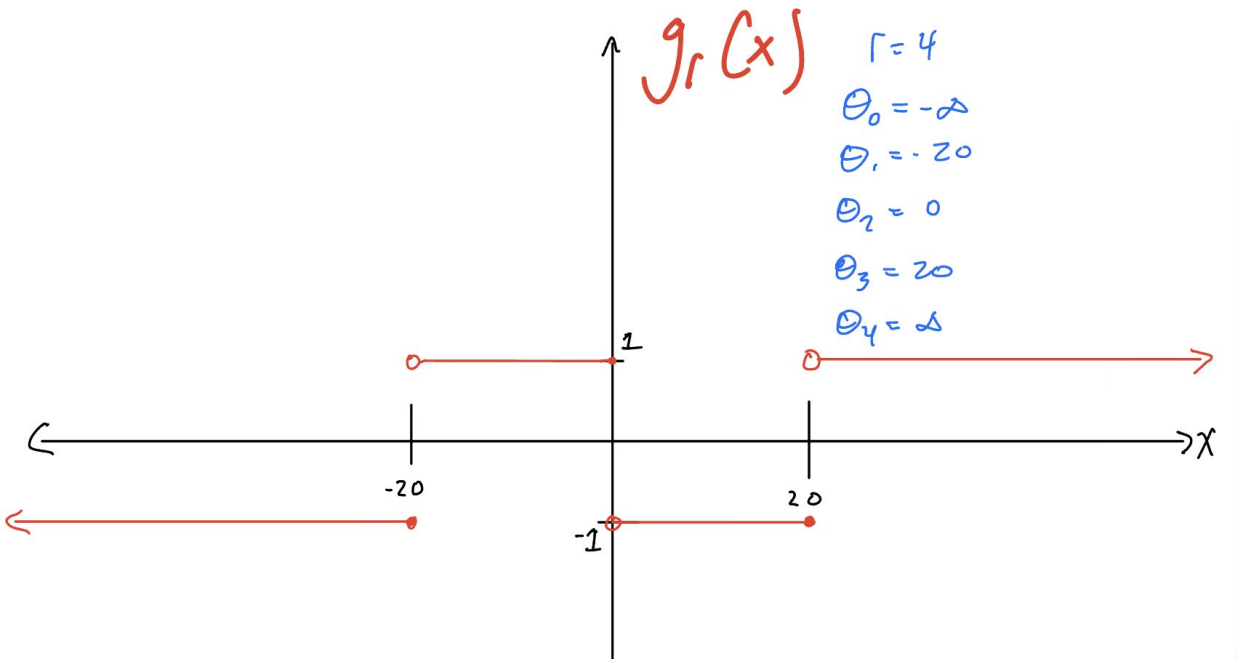
since $x \in (0, 20]$ implies $x \notin (\theta_{j-1}, \theta_j]$ for all $j \neq 3$ by (57).
Finally, for all $x \in (20, \infty]$ (i.e. all x we have not yet discussed), we have

$$g_r(x) = \alpha_4 I\{x \in (20, \infty]\} + \sum_{j=1}^{3} \alpha_j I\{x \in (\theta_{j-1}, \theta_j]\} = \alpha_4 I\{x \in (20, \infty]\} = \alpha_4 = (-1)^4 = 1 \qquad (61)$$

Thus, for $r = 4$ and $\theta_0 < \theta_1 < \theta_2 < \theta_3 < \theta_4$ as defined in (56), we have

$$g_r(x) = \begin{cases} -1 \text{ if } x \in (-\infty, -20] \cup (0, 20] \\ 1 \text{ if } x \in (-20, 0] \cup (20, \infty) \end{cases} \qquad (62)$$

We can now plot g_r as a piece-wise function of x:



which completes part (a).

(b) We want to show that, for all possible g_r,

$$g_r(x) = sign(g(x)) \qquad (63)$$

for some $g \in G$. To do so, it suffices to show:

(i) $h(x) = sign(g(x))$ for some $g \in G$.

(ii) $h(x) = g_r(x)$.

First, we will show (i). To do so, we will construct a valid $g \in G$, then show that $h(x) = sign(g)$. Let

$$\alpha_1 = \frac{1}{2} \text{ and } \alpha_2 = \cdots = \alpha_r = 1 \qquad (64)$$

Next, define

$$f_i(x) = f_{\theta_{i-1}, sign(w_i)}(x) := sign(w_i) \cdot sign(x - \theta_{i-1}) \qquad (65)$$

for all $i \in \{1, ..., r\}$. Since $sign(w_i) \in \{-1, +1\}$, we have

$$f_i(x) = sign(x - \theta_{i-1}) \cdot b \in \mathcal{F} \qquad (66)$$

for all $i \in \{1, ..., r\}$. Also, by (64), we have $\alpha_i \geq 0$ for all $i \in \{1, ..., r\}$. Thus, for all $r \geq 1$, by the definition of G, we have

$$g(x) := \sum_{i=1}^{r} \alpha_i f_i \in G \qquad (67)$$

Now, we just have to show that $h(x) = sign(g(x))$. Note that

$$\begin{aligned} g(x) := \quad & \sum_{i=1}^{r} \alpha_i f_i = \alpha_1 f_1 + \sum_{i=2}^{r} \alpha_i f_i = \frac{1}{2} sign(-1) sign(x - \theta_0) + \sum_{i=2}^{r} sign(w_i) sign(x - \theta_{i-1}) \\ = \quad & -\frac{1}{2} sign(x - \theta_0) + \sum_{i=2}^{r} sign((-1)^i) sign(x - \theta_{i-1}) \qquad (68) \end{aligned}$$

Since $sign((-1)^i) = (-1)^i$ for all $i \geq 1$, we have

$$g(x) = -\frac{1}{2} sign(x - \theta_0) + \sum_{i=2}^{r} (-1)^j sign(x - \theta_{i-1}) \qquad (69)$$

Noting that $-\frac{1}{2} =: w_1$ and $(-1)^i =: w_i$ for all $i > 1$, we can rewrite $g(x)$ as

$$g(x) = w_1 sign(x - \theta_0) + \sum_{i=2}^{r} w_i sign(x - \theta_{i-1}) = \sum_{i=1}^{r} w_i sign(x - \theta_{i-1}) \qquad (70)$$

Combining (70) with the definition of $h(x)$ from the problem statement, we find

$$h(x) := sign(\sum_{i=1}^{r} w_i sign(x - \theta_{i-1})) = sign(g(x)) \qquad (71)$$

Since $g \in G$ by (67), this completes the proof that

$$h(x) = sign(g(x)) \text{ for some } g \in G \qquad (72)$$

Now, we just need to show that $g_r(x) = h(x)$. We will consider several different cases. First, fix $x \in \mathbb{R}$, and define $k \in \{1, ..., r\}$ to be the index such that $x \in (\theta_{k-1}, \theta_k]$. Note that, by (57), this definition implies

$$x \notin (\theta_{j-1}, \theta_j] \text{ for all } j \neq k \qquad (73)$$

Also, since $\theta_{k-1} < x \leq \theta_k$, we know

$$sign(x - \theta_i) = +1 \text{ for all } i \in \{1, ..., k-1\} \qquad (74)$$

and

$$sign(x - \theta_i) = -1 \text{ for all } i \in \{k, ..., r\} \qquad (75)$$

Note that, since $x = k$ is possible by the definition of k, we are defining $sign(0) = -1$ for (75) to hold. The sign of 0 is arbitrary, so it is fine to do this. Also, whenever $x \neq k$, $x < k$, and $sign(x - \theta_k) := -1$. Thus, defining $sign(0) = -1$ makes sense to maintain the consistency of the term $sign(x - \theta_k)$.
Since $k \in \mathbb{Z}$, k is either even or odd. We consider these cases separately.

Case 1: First, we consider the case when k is odd. By (73), we can easily compute that

$$g_r(x) = \alpha_k I\{x \in (\theta_{k-1}, \theta_k]\} + \sum_{j \neq k, j \in \{1,...,r\}} \alpha_j I\{x \in (\theta_{j-1}, \theta_j]\} = \alpha_k = (-1)^k = -1 \qquad (76)$$

with the last equality following from the assumption that k is odd.
Claim: We claim that, for all odd $m \in \mathbb{Z}$ where $1 \leq m \leq k$,

$$\sum_{i=1}^{m} w_i sign(x - \theta_{i-1}) = -\frac{1}{2}$$

Proof: We induct on m.
Base Case: $m = 1$. We can easily compute

$$\sum_{i=1}^{1} w_i sign(x - \theta_0) = w_1 \cdot (1) = -\frac{1}{2}$$

so the claim holds for the base case.
Inductive Hypothesis: Assume that

$$\sum_{i=1}^{m} w_i sign(x - \theta_{i-1}) = -\frac{1}{2}$$

for all odd integers m such that $1 \leq m \leq l \leq k-2$ for some odd integer l.
Inductive Step: Consider $m = l + 2$. Note that $l + 2$ is odd since l is odd. With the help of (74), we can easily compute that

$$\begin{aligned} \sum_{i=1}^{l+2} w_i sign(x - \theta_{i-1}) = & \quad \sum_{i=1}^{l} w_i sign(x - \theta_{i-1}) + w_{l+1} sign(x - \theta_l) + w_{l+2} sign(x - \theta_{l+1}) \\ = & \quad \sum_{i=1}^{l} w_i sign(x - \theta_{i-1}) + (-1)^{l+1}(1) + (-1)^{l+2}(1) \qquad (77) \end{aligned}$$

Since l is odd, $l + 1$ is even, while $l + 2$ is odd, so

$$(-1)^{l+1} = 1 \quad \text{while} \quad (-1)^{l+2} = -1 \qquad (78)$$

Also, by the inductive hypothesis, we have

$$\sum_{i=1}^{l} w_i sign(x - \theta_{i-1}) = -\frac{1}{2} \qquad (79)$$

Plugging (79) and (78) into (77) yields

$$\sum_{i=1}^{l+1} w_i sign(x-\theta_{i-1}) = -\frac{1}{2} + 1(1) + (-1)(1) = -\frac{1}{2} + 1 - 1 = -\frac{1}{2} \qquad (80)$$

The conclusion that

$$\sum_{i=1}^{m} w_i sign(x-\theta_{i-1}) = -\frac{1}{2} \qquad (81)$$

follows from (80) by induction for all odd integers m such that $1 \leq m \leq k$. Thus, when k is odd, we have

$$\sum_{i=1}^{k} w_i sign(x-\theta_{i-1}) = -\frac{1}{2} \qquad (82)$$

The value of $\sum_{i=k+1}^{r} w_i sign(x-\theta_{i-1})$ depends on whether r is even or odd, so we split into two more sub-cases:

Case 1.1: k is odd AND r is even.
Claim: We claim that, for all even integers m such that $k+1 \leq m \leq r$, we have

$$\sum_{i=k+1}^{m} w_i sign(x-\theta_{i-1}) = -1$$

Proof: We induct on m.
Base Case: Since k is odd, $k+1$ is even, so we start with $m = k+1$. With the help of (75), we can directly compute

$$\sum_{i=k+1}^{k+1} w_i sign(x-\theta_{i-1}) = w_{k+1} sign(x-\theta_k) = (-1)^{k+1} sign(x-\theta_k) = 1(-1) = -1$$

with the last equality following since k is odd, so $k+1$ is even. Thus, the claim holds for the base case.
Inductive Hypothesis: Assume that, for all even integers m such that $k+1 \leq m \leq l \leq r-2$, where l is an even integer, we have

$$\sum_{i=k+1}^{m} w_i sign(x-\theta_{i-1}) = -1$$

Inductive Step: Consider $m = l+2$. Since l is even, $l+2$ is even. Combining (75) with the inductive hypothesis, we can compute that

$$\begin{aligned} \sum_{i=k+1}^{l+2} w_i sign(x-\theta_{i-1}) = & \sum_{i=k+1}^{l} w_i sign(x-\theta_{i-1}) + w_{l+1} sign(x-\theta_l) + w_{l+2} sign(x-\theta_{l+1}) \\ = & -1 + (-1)^{l+1}(-1) + (-1)^{l+2}(-1) \qquad (83) \end{aligned}$$

Since l is even, $l+1$ is odd, while $l+2$ is even, so

$$(-1)^{l+1} = -1 \quad \text{and} \quad (-1)^{l+2} = 1 \qquad (84)$$

Plugging the results from (84) into (83) yields

$$\sum_{i=k+1}^{l+2} w_i sign(x-\theta_{i-1}) = -1 + (-1)(-1) + 1(-1) = -1 + 1 - 1 = -1 \qquad (85)$$

The conclusion that

$$\sum i = k + 1^m w_i sign(x - \theta i - 1) = -1 \qquad (86)$$

follows for all odd integers m such that $k+1 \leq m \leq r$ from (86) by induction. Thus, we have

$$\sum_{i=k+1}^{r} w_i sign(x-\theta_{i-1}) = -1 \qquad (87)$$

whenever k is odd AND r is even. Combining (87) with (82), we find

$$\sum_{i=1}^{r} w_i sign(x-\theta_{i-1}) = \sum_{i=1}^{k} w_i sign(x-\theta i-1) + \sum_{i=k+1}^{r} w_i sign(x-\theta i-1) = -\frac{1}{2} - 1 = -\frac{3}{2} \qquad (88)$$

Thus, by the definition of $h(x)$ from the problem statement, we have

$$h(x) := sign(\sum_{i=1}^{r} w_i sign(x-\theta_{i-1})) = sign(-\frac{3}{2}) = -1 \qquad (89)$$

Combining (89) and (76), we find that

$$h(x) = -1 = g_r(x) \qquad (90)$$

for all odd k and even r, which completes *Case 1.1.*

Case 1.2: k is odd AND r is odd.
Claim: For all odd integers m such that $k+1 < m \leq r$, we have

$$\sum_{i=k+1}^{m} w_i sign(x-\theta_{i-1}) = 0$$

Proof: We induct on m.
Base Case: Note that, since k is odd, $k+1$ is even, so $k+2$ is the first odd integer greater than $k+1$. Thus, our base case is $m = k+2$. Using (75), we can directly compute

$$\begin{aligned} \sum_{i=k+1}^{k+2} w_i sign(x-\theta_{i-1}) = &\quad w_{k+1} sign(x-\theta_k) + w_{k+2} sign(x-\theta_{k+1}) \\ = &\quad (-1)^{k+1}(-1) + (-1)^{k+2}(-1) = 1(-1) + (-1)(-1) = -1+1 = 0 \end{aligned} \qquad (91)$$

so the claim holds for the base case.
Inductive Hypothesis: Assume that, for all odd integers m such that $k+1 < m \leq l \leq r-2$ for some odd integer l, we have

$$\sum_{i=k+1}^{m} w_i \cdot sign(x-\theta_{i-1}) = 0$$

Inductive Step: Consider $m = l+2$. Combining the inductive hypothesis with (75), we find

$$\begin{aligned} \sum_{i=k+1}^{l+2} w_i sign(x-\theta_{i-1}) = &\quad \sum_{i=k+1}^{l} w_i sign(x-\theta_{i-1}) + w_{l+1} sign(x-\theta_l) + w_{l+2} sign(x-\theta_{l+1}) \\ = &\quad 0 + (-1)^{l+1}(-1) + (-1)^{l+2}(-1) \end{aligned} \qquad (92)$$

Since l is odd, $l+1$ is even, while $l+2$ is odd, so

$$(-1)^{l+1} = 1 \quad \text{and} \quad (-1)^{l+2} = -1 \qquad (93)$$

Plugging (93) into (92) yields

$$\sum_{i=k+1}^{l+2} w_i sign(x-\theta_{i-1}) = 1(-1)+(-1)(-1) = -1+1 = 0 \qquad (94)$$

The conclusion that

$$\sum_{i=k+1}^{m} w_i sign(x-\theta_{i-1}) = 0 \qquad (95)$$

follows for all odd integers m such that $k+1 < m \leq r$ from (94) by induction. Thus, since r is odd, we have

$$\sum_{i=k+1}^{r} w_i sign(x-\theta i-1) = 0 \qquad (96)$$

whenever k and r are both odd.
Combining (96) with (82), we find

$$\sum_{i=1}^{r} w_i sign(x-\theta i-1) = \sum_{i=1}^{k} w_i sign(x-\theta i-1) + \sum_{i=k+1}^{r} w_i sign(x-\theta i-1) = -\frac{1}{2}+0 = -\frac{1}{2} \qquad (97)$$

Applying the definition of $h(x)$ from the problem statement to (97) yields

$$h(x) := sign(\sum_{i=1}^{r} w_i sign(x-\theta_{i-1}) = sign(-\frac{1}{2}) = -1 \qquad (98)$$

whenever k and r are both odd. Combining (98) with (76), we find

$$h(x) = -1 = g_r(x) \qquad (99)$$

whenever k and r are both odd, which completes *Case 1.2.*
Combining (99) with (90), we find that

$$h(x) = -1 = g_r(x) \qquad (100)$$

whenever k is odd for all $r \geq 1$, which completes all of *Case 1.*

Case 2: Now, consider the possibility that k is even. By (73), we can easily compute that

$$g_r(x) = \alpha_k I\{x \in (\theta_{k-1}, \theta_k]\} + \sum_{j \neq k, j \in \{1,\ldots,r\}} \alpha_j I\{x \in (\theta_{j-1}, \theta_j]\} = \alpha_k = (-1)^k = 1 \qquad (101)$$

with the last equality following from the assumption that k is even.
Claim: We claim that, for all even integers m such that $1 < m \leq k$, we have

$$\sum_{i=1}^{m} w_i sign(x-\theta_{i-1}) = \frac{1}{2}$$

Proof: We induct on m.
Base Case: The smallest such m is $m = 2$. Using (74), we can directly compute that

$$\sum_{i=1}^{2} w_i sign(x-\theta_{i-1}) = w_1 sign(x-\theta_0) + w_2 sign(x-\theta_1) = -\frac{1}{2}(1)+(-1)^2(1) = -\frac{1}{2}+1 = \frac{1}{2}$$

so the claim holds for the base case.
Inductive Hypothesis: Assume that, for all even integers such that $1 < m \leq l \leq k-2$ for some even integer l, we have

$$\sum_{i=1}^{m} w_i sign(x - \theta_{i-1}) = \frac{1}{2}$$

Inductive Step: Consider $m = l + 2$. Combining (74) with the inductive hypothesis, we find

$$\begin{aligned} \sum_{i=1}^{l+2} w_i sign(x - \theta_{i-1}) = & \quad \sum_{i=1}^{l} w_i sign(x - \theta_{i-1}) + w_{l+1} sign(x - \theta_l) + w_{l+2} sign(x - \theta_{l+1}) \\ = & \quad \frac{1}{2} + (-1)^{l+1}(1) + (-1)^{l+2}(1) \qquad (102) \end{aligned}$$

Since l is even, $l + 1$ is odd, while $l + 2$ is even, so

$$(-1)^{l+1} = -1 \quad \text{and} \quad (-1)^{l+2} = 1 \qquad (103)$$

Combining (102) with (103) yields

$$\sum_{i=1}^{l+2} w_i sign(x - \theta_{i-1}) = \frac{1}{2} + -1(1) + 1(1) = \frac{1}{2} - 1 + 1 = \frac{1}{2} \qquad (104)$$

The conclusion that

$$\sum_{i=1}^{m} w_i sign(x - \theta_{i-1}) = \frac{1}{2} \qquad (105)$$

follows from (104) by induction for all even integers m such that $1 < m \leq k$ whenever k is even. The conclusion in (105) directly implies that

$$\sum_{i=1}^{k} w_i sign(x - \theta_{i-1}) = \frac{1}{2} \qquad (106)$$

for all even k.
Once more, the value of $\sum_{i=k+1}^{r} w_i sign(x - \theta_{i-1})$ depends on whether r is even or odd, so we split into two more sub-cases:

Case 2.1: k is even AND r is even.
Claim: We claim that, for all even integers m such that $k + 1 < m \leq r$, we have

$$\sum_{i=k+1}^{m} w_i sign(x - \theta_{i-1}) = 0$$

Proof: We induct on m.
Base Case: The smallest such even integer m is $m = k + 2$ since k is even. Using (75) alongside the fact that $k + 1$ is odd and $k + 2$ is even, we can directly compute that

$$\begin{aligned} \sum_{i=k+1}^{k+2} w_i sign(x - \theta_{i-1}) = & \quad w_{k+1} sign(x - \theta_k) + w_{k+2} sign(x - \theta_{k+1}) \\ = & \quad (-1)^{k+1}(-1) + (-1)^{k+2}(-1) = (-1)(-1) + 1(-1) = 1 - 1 = 0 \end{aligned}$$

so the claim holds for the base case.
Inductive Hypothesis: Assume that, for all even integers m such that $k + 1 < m \leq l \leq r - 2$ for some even integer l, we have

$$\sum_{i=k+1}^{m} w_i sign(x - \theta_{i-1}) = 0$$

Inductive Step: Consider $m = l + 2$. Combining (75) with the inductive hypothesis, we find

$$\begin{aligned}\sum_{i=k+1}^{l+2} w_i sign(x-\theta_{i-1}) = & \sum_{i=k+1}^{l} w_i sign(x-\theta_{i-1}) + w_{l+1} sign(x-\theta_l) + w_{l+2} sign(x-\theta_{l+1}) \\ = & \; 0 + (-1)^{l+1}(-1) + (-1)^{l+2}(-1) = (-1)^{l+1}(-1) + (-1)^{l+2}(-1) \end{aligned} \tag{107}$$

Since l is even, $l+1$ is odd, while $l+2$ is even, so

$$(-1)^{l+1} = -1 \quad \text{and} \quad (-1)^{l+2} = 1 \tag{108}$$

Plugging the results from (108) into (107) yields

$$\sum_{i=k+1}^{l+2} w_i sign(x-\theta_{i-1}) = (-1)(-1) + 1(-1) = 1 - 1 = 0 \tag{109}$$

The conclusion that

$$\sum_{i=k+1}^{m} w_i sign(x-\theta_{i-1}) = 0 \tag{110}$$

follows from (109) by induction for all even integers m such that $k+1 < m \leq r$, whenever k is even. The conclusion in (110) directly implies that

$$\sum_{i=k+1}^{r} w_i sign(x-\theta_{i-1}) = 0 \tag{111}$$

whenever k is even and r is even. Combining (111) and (106) yields

$$\sum_{i=1}^{r} w_i sign(x-\theta_{i-1}) = \sum_{i=1}^{k} w_i sign(x-\theta_{i-1}) + \sum_{i=k+1}^{r} w_i sign(x-\theta_{i-1}) = \frac{1}{2} + 0 = \frac{1}{2} \tag{112}$$

for all even k and even r. Applying the definition of $h(x)$ from the problem statement to (112) yields

$$h(x) := sign(\sum_{i=1}^{r} w_i sign(x-\theta_{i-1})) = sign(\frac{1}{2}) = 1 \tag{113}$$

for all even k and even r. Comparing (113) with (101), we see that

$$h(x) = 1 = g_r(x) \tag{114}$$

for all even k and even r, which completes *Case 2.1.*

Case 2.2: k is even AND r is odd.
Claim: We claim that, for all odd integers m such that $k+1 \leq m \leq r$, we have

$$\sum_{i=k+1}^{m} w_i sign(x-\theta_{i-1}) = 1$$

Proof. We induct on m.
Base Case: Since k is even, the smallest such odd integer m is $m = k+1$. Combining (75) with the fact that $k+1$ is odd, we can directly compute that

$$\sum_{i=k+1}^{k+1} w_i sign(x-\theta_{i-1}) = w_{k+1} sign(x-\theta_k) = (-1)^{k+1}(-1) = (-1)(-1) = 1$$

so the claim holds for the base case.
Inductive Hypothesis: Assume that, for all odd integers m such that $k+1 \leq m \leq l \leq r-2$, for some odd integer l, we have

$$\sum_{i=k+1}^{m} w_i(x-\theta_{i-1}) = 1$$

Inductive Step: Consider $m = l+2$. Combining the inductive hypothesis with (75), we find

$$\begin{aligned}\sum_{i=k+1}^{l+2} w_i sign(x-\theta_{i-1}) = & \sum_{i=k+1}^{l} w_i sign(x-\theta_{i-1}) + w_{l+1} sign(x-\theta_l) + w_{l+2} sign(x-\theta_{l+1}) \\ = & \quad 1 + (-1)^{l+1}(-1) + (-1)^{l+2}(-1) \qquad (115)\end{aligned}$$

Since l is odd, $l+1$ is even, while $l+2$ is odd, so

$$(-1)^{l+1} = 1 \quad \text{and} \quad (-1)^{l+2} = -1 \qquad (116)$$

Plugging the results from (116) into (115), we find

$$\sum_{i=k+1}^{l+2} w_i sign(x-\theta_{i-1}) = 1 + 1(-1) + -1(-1) = 1 - 1 + 1 = 1 \qquad (117)$$

The conclusion that

$$\sum_{i=k+1}^{m} w_i sign(x-\theta_{i-1}) = 1 \qquad (118)$$

follows from (117) by induction for all odd integers m such that $k+1 \leq m \leq r$, whenever k is even and r is odd. That means, for even k and odd r, we have

$$\sum_{i=k+1}^{r} w_i sign(x-\theta_{i-1}) = 1 \qquad (119)$$

Combining (119) with (106), we find

$$\sum_{i=1}^{r} w_i sign(x-\theta_{i-1}) = \sum_{i=1}^{k} w_i sign(x-\theta_{i-1}) + \sum_{i=k+1}^{r} w_i sign(x-\theta_{i-1}) = \frac{1}{2} + 1 = \frac{3}{2} \qquad (120)$$

whenever k is even and r is odd. Applying the definition of $h(x)$ from the problem statement to (120) yields

$$h(x) = sign(\sum_{i=1}^{r} w_i sign(x-\theta_{i-1})) = sign(\frac{3}{2}) = 1 \qquad (121)$$

for all even k and odd r. Comparing (121) with (101), we find

$$h(x) = 1 = g_r(x) \qquad (122)$$

for all even k and odd r, which completes *Case 2.2.*
Combining (122) with (114), we find that

$$h(x) = 1 = g_r(x) \qquad (123)$$

for all even k and for all $r \geq 1$, which completes the entirety of *Case 2.*
Combining (123) with (100), we find

$$h(x) = g_r(x) \qquad (124)$$

for all k (i.e. both all even and all odd k) and all $r \geq 1$, which completes the proof of part (ii). Combining (124) with (72), we conclude that any function g_r can be realized as an element of the class G after taking the sign. Specifically, we conclude that, for any g_r, g_r satisfies

$$g_r(x) = sign(g(x)) \qquad (125)$$

for some $g \in G$, which completes part (b).

MATH 407: Probability Theory

All assignments in this section were written by Joshua P. Swanson, RTPC Assistant Professor of Mathematics, USC. Solutions to assignments 1 through 14 are provided.

Assignment 1

Math 407 (Swanson) – Spring 2023
Homework 1
Due Friday 1/13, 11:59pm

Name: Emerson Kahle Section: 39981

- You must upload your solutions to Gradescope as **one single, high-quality PDF**. You can convert paper-based work to a high-quality PDF using a scanning app for mobile devices, such as Adobe Scan (free, available for iOS and Android, can do multiple pages) or many others. If necessary, you can combine or merge multiple PDF's into a single PDF using a variety of services, such as Adobe Acrobat's cloud-based merge tool.
- After you upload, you must match each question with its corresponding page using Gradescope's interface. This allows graders to spend more time giving you feedback instead of hunting through submissions.
- Answers without supporting work will receive no credit. Show your work.
- You are encouraged to work together on homework, but **you must write up your solutions separately in your own words**. Copying from your fellow students or other sources is a serious academic integrity violation. In particular, you may not use "tutoring" services which simply provide answers.
- You are encouraged to typeset your solutions in LaTeX. Source code has been provided on Blackboard. Overleaf is a popular cloud-based editor.
- Problem numbers refer to the course textbook, though the problems may have been modified significantly.

1. (Ross P1.6) A well-known nursery rhyme starts as follows:

> "As I was going to St. Ives
> I met a man with 7 wives.
> Each wife had 7 sacks.
> Each sack had 7 cats.
> Each cat had 7 kittens..."

How many kittens did the traveler meet?
How many were going to St. Ives?

Solution.
To count the total number of kittens that the traveler met, we operate under the assumption that the traveler meets all of the wives, sacks, cats, and kittens that the man has. Since the traveler describes in first person the specific quantities of wives, sacks, cats, and kittens in the man's possession, it is reasonable to assume that the traveler actually met these people, animals, and objects. Furthermore, since the only specific quantities described by the traveler relate to the numbers of wives, sacks, cats and kittens that the man has, this assumption is necessary to conclude that the traveler met a specific number of kittens.
Now, we can make a series of observations to determine the specific number of kittens that the traveler met (under the above assumption).
Observation 1:
Since each sack has 7 cats, each with 7 kittens, each sack has exactly $7 * 7 = 49$ kittens.
Observation 2:
Since each wife has 7 sacks, each with 49 kittens, each wife has $7 * 49 = 343$ kittens.
Observation 3:
Since the man had 7 wives, each with 343 kittens, the man's wives combine to have exactly $7 * 343 = 2401$ kittens.
Observation 4:
Since the traveler met the man, and we assume he met all of the man's wives, sacks, cats, and kittens, we know the traveler met exactly 2401 kittens.

To answer the second question, we again must make an assumption.
Since the traveler meets the man on the way to a destination, it is most likely that the man and his wives, sacks, cats, and kittens are moving in the opposite direction of the traveler.
Since the traveler is on his way towards St. Ives, this implies that the man and his wives, sacks, cats, and kittens are heading away from St. Ives. Thus, we assume that the man and his wives, sacks, cats, and kittens are all moving away from St. Ives.
If the man and traveler were both traveling to St. Ives, it seems more likely they would meet at their shared destination than at some point along the way.
Under this assumption, it is clear that, of all the people, objects, and animals described by the traveler, only the traveler himself is going to St. Ives.
Thus, the answer to "how many were going to St. Ives" is exactly 1, the traveler himself.
Note: Since the second question does not specify kittens, we assume that it refers to all people, objects, and animals described by the traveler, including himself.

2. (Ross P1.1)

   (a) How many different 7-plate license plates are possible if the first 2 places are for letters and the other 5 are for numbers?

   (b) Repeat part (a) under the assumption that no letter or number can be repeated in a single license plate.

Solution.
(a)
Since the letters can be repeated, and there are 26 letters, there are $26 * 26 = 676$ distinct possibilities for the first two places.
Since the numbers can be repeated, and there are 10 digits, there are $10 * 10 * 10 * 10 * 10 = 10^5 = 100,000$ possibilities for the last five places.
Since each one of the 676 possibilities for the first two places can be paired with any of the 100,000 possibilities for the last five places to produce a distinct 7-plate license plate, there are exactly $676 * 100,000 = 67,600,000$ different 7-plate license plates, if the first two places are for letters and the other five are for numbers.
(b)
Since the letters cannot be repeated, there are 26 options for the first letter but only 25 options for the second, meaning there are exactly $26 * 25 = 650$ distinct possibilities for the first two places.
Since the numbers cannot be repeated, there are 10 options for the first digit, but only 9 options for the second, then 8 for the third, 7 for the fourth, and 6 for the fifth.
Thus, there are exactly $10 * 9 * 8 * 7 * 6 = 30,240$ distinct possibilities for the last five places.
Just like in part (a), each one of the 650 possibilities for the first two places can be paired with any of the 30,240 possibilities for the last five places to produce a distinct 7-plate license plate. Thus, there are exactly $650 * 30,240 = 19,656,000$ different 7-plate license plates, if the first two places are for letters and the other five are for numbers, and neither numbers nor letters can be repeated in a single license plate.

3. (Ross TE1.3) In how many ways can r objects be selected from a set of n objects if the order of selection is considered relevant?

Solution.
To select r objects from an n object set, we must first select 1 object from the set, for which there are exactly n possibilities.
For the second object, we now have exactly $n-1$ possibilities, so there are $n*(n-1) = n^2 - n$ distinct possibilities for the first two objects that we select (assuming $r \geq 2$).
This pattern continues until we have selected r objects, at which point we will have $n-r$ objects in the set that have yet to be selected.
This implies that we had $n-r+1$ options when we selected the r'th object.
Therefore, we can select r distinct objects from a set of n objects in exactly $n(n-1)...(n-r+1)$ ways if the order of selection is considered relevant.
Multiply this expression by $1 = \frac{(n-r)!}{(n-r)!}$ to obtain:

$$n(n-1)...(n-r+1)\frac{(n-r)!}{(n-r)!} = \frac{n(n-1)...(n-r+1)(n-r)!}{(n-r)!} = \frac{n!}{(n-r)!}$$

Thus, there are exactly $\frac{n!}{(n-r)!}$ ways that r distinct objects can be selected from a set of n objects, if the order of selection is considered relevant.
Note: Although the combinatorial method used to obtain this answer relies on $r \geq 2$, the final formula works even if $r = 1$, at which point $\frac{n!}{(n-r)!} = \frac{n!}{(n-1)!} = n$, as expected.

4. (Ross P1.15) Consider a group of 20 people. If everyone shakes hands with everyone else, how many handshakes take place?

   Prove that
   $$1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$
   by interpreting each side in terms of handshakes.

   Solution.
   The first person shakes hands with 19 distinct people.
   The second person adds 18 new handshakes by shaking hands with all 18 people besides the first person.
   This pattern continues until the nineteenth person only adds 1 new handshake by shaking hands with the twentieth person, who adds 0 new handshakes.
   Thus, a total of $1 + 2 + ... + 18 + 19 = 190$ distinct handshakes take place.

   Proof.
   Generalizing the above solution, it is clear that the total number of handshakes in a similar group of $n + 1$ people is the left side of the equation $(1 + 2 + ... + n)$.
   Now, let's count the total number of handshakes among a similar group of $n + 1$ people in a different way.
   Each of the $n + 1$ people shakes hands with each of the n other people, for a total of $n(n + 1)$ handshakes.
   We want to consider only distinct handshakes, and $n(n + 1)$ counts each handshake twice (i.e. counts both the handshake between the first person and second person and the handshake between the second person and first person, even though they are not distinct), so we must divide our product by 2.
   Thus, the total number of distinct handshakes is $\frac{n(n+1)}{2}$.
   Since we already know the total number of distinct handshakes is $1 + 2 + ... + n$, we have proven that
   $$1 + 2 + ... + n = \frac{n(n+1)}{2}$$

5. (Ross TE1.8) Prove that

$$\binom{n+m}{r} = \binom{n}{0}\binom{m}{r} + \binom{n}{1}\binom{m}{r-1} + \cdots + \binom{n}{r}\binom{m}{0}$$

Hint: Consider a group of n men and m women. How many groups of size r are possible?

Proof.
Consider a group of n men and m women, from which we want to select a subgroup of size r, where $0 \leq r \leq m+n$. By definition, there are $\binom{n+m}{r}$ ways to do this.
In any such subgroup, some number, i, of the r people will be from the group of n men, where $0 \leq i \leq r$.
The remaining $r-i$ people from the r person subgroup must be from the group of m women.
Since $0 \leq i \leq r$, $0 \leq r-i \leq r$.
Also, for each i, there are exactly $\binom{n}{i}\binom{m}{r-i}$ groups of size r.
Since i can be any integer value from 0 to n, there are exactly

$$\sum_{i=0}^{r} \binom{n}{r}\binom{m}{r-i} = \binom{n}{0}\binom{m}{r} + \binom{n}{1}\binom{m}{r-1} + \cdots + \binom{n}{r}\binom{m}{0}$$

ways to select a subgroup of r people from a group of n men and m women.
Therefore,

$$\binom{n+m}{r} = \binom{n}{0}\binom{m}{r} + \binom{n}{1}\binom{m}{r-1} + \cdots + \binom{n}{r}\binom{m}{0}$$

which concludes the proof.

6. (Ross P1.8) When all letters are used, how many different letter arrangements can be made from the letters

   (a) FLUKE

   (b) PROPOSE

   (c) MISSISSIPPI

   Solution.

   (a)
   FLUKE consists of 5 distinct letters, so their are 5! distinct ways to arrange the letters in FLUKE.
   Thus, the number of different letter arrangements when all letters in FLUKE are used once is 5!.

   (b)
   There are 7 letters in PROPOSE, but P and O both appear twice.
   Number them P_1, P_2, O_1, O_2. Now, $P_1RO_1P_2O_2SE$ consists of 7 distinct letters.
   Therefore, there are 7! distinct ways to arrange all the letters in $P_1RO_1P_2O_2SE$.
   Since 7! considers $P_1RO_1P_2O_2SE$, $P_2RO_1P_1O_2SE$, $P_1RO_2P_2O_1SE$, and $P_2RO_2P_1O_1SE$
   all to be distinct, it quadruple counts each of the possible letter arrangments for PROPOSE.
   Therefore, there are $\frac{7!}{4} = 1260$ distinct letter arrangments when all letters in PROPOSE
   are used once.

   (c)
   There are 11 letters in MISSISSIPPI, but I and S appear 4 times each, while P appears twice.
   Similar to part (b), if we number the I's, S's, and P's to distinguish them, then we could create
   11! distinct letter arrangments using each letter in $MI_1S_1S_2I_2S_3S_4I_3P_1P_2I_4$.
   However, each distinct letter arrangment that uses each letter in MISSISSIPPI once is
   overcounted $4! * 4! * 2!$ times by 11!, due to the 4! permutations of the 4 I's, the 4!
   permutations of the 4 S's, and the 2! permutations of the 2 P's.
   Thus, the total number of distinct letter arrangments that use each letter in MISSISSIPPI once is

$$\frac{11!}{(4!)^2 * 2!} = 34,650$$

7. (Ross TE1.12) Consider the following combinatorial identity:

$$\sum_{k=1}^{n} k\binom{n}{k} = n \cdot 2^{n-1}.$$

(a) Present a combinatorial argument for this identity by considering a set of n people and determining, in two ways, the number of possible selections of a committee of any size and a chairperson for the committee.

Hint:

(i) How many possible selections are there of a committee of size k and its chairperson?

(ii) How many possible selections are there of a chairperson and the other committee members?

(b) Now present an algebraic argument for this identity.

Hint: Differentiate the binomial theorem.

Proof.
(a)
We want to count the total number of ways to select a committee of size k and a chairperson for that committee from a set of n people.
Note: the number of ways to make these selections is the same regardless of whether the committee or chairperson is chosen first.
First, consider choosing the committee first, then the chairperson from that committee. The committee must have at least one person (the chairperson), and it can be any size, so it can include all people from the set. Therefore, the size of the committee, k, can be any element of $\{1, 2, ..., n\}$. For any such $k \in \{1, 2, ..., n\}$, there are $\binom{n}{k}$ ways to choose a group of k people to form the committee from the set of n people. Since the chairperson must be a member, this leaves exactly k choices for the chairperson. Therefore, for any such k, there are exactly $k\binom{n}{k}$ ways to choose a committee of size k then a chairperson from that committee. Since k can be any integer value from 1 to n, we can compute that the total number of ways to choose a committee of any size then a chairperson from that committee is

$$\sum_{k=1}^{n} k\binom{n}{k} \tag{3}$$

which is the left side of the identity.
Now, we can count this same quantity, but first we will choose the chairperson, then we will choose the committee around the chairperson. Since there are n people in the set, we have exactly n options for the chairperson. Since the committee can be any size from 1 to n, and the chairperson is the first member of the committee, the size of the committee excluding the chairperson, which we'll call j, can be any integer from 0 to $n-1$. For any such j, there are exactly $\binom{n-1}{j}$ ways to choose the members of the of set that will join the chairperson in the committee. Summing over all possible values of j, we find that the total number of ways to select a chairperson then form the committee around that chairperson is

$$n\sum_{j=0}^{n-1} \binom{n-1}{j} \tag{4}$$

$\sum_{j=0}^{n-1} \binom{n-1}{j}$ counts every way to choose 0 elements from an $n-1$ element set, every way to choose 1 element from an $n-1$ element set, and so forth, all the way until every way to choose $n-1$ elements from an $n-1$ element set. Therefore, $\sum_{j=0}^{n-1} \binom{n-1}{j}$ = the number of subsets of a set of size $n-1$. To construct a subset of a set of size $n-1$, we can either choose to include or exclude each element of the set. All of these choices are independent, which yields 2^{n-1} total possible subsets (multiplying the number of possible subsets by 2 for each additional element).

Thus, $\sum_{j=0}^{n-1} \binom{n-1}{j} = 2^{n-1}$. Plugging this into (2), we find that the total number of ways to select a chairperson then form the committee around that chairperson is

$$n2^{n-1} \tag{5}$$

Since (1) and (3) both describe the number of possibile selections for the same situation, they are equivalent, which concludes the combinatorial proof that

$$\sum_{k=1}^{n} k\binom{n}{k} = n2^{n-1}$$

(b)
Newton's Binomial Theorem specifies that

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \tag{6}$$

Plugging $y = 1$ into (4) yields:

$$(x+1)^n = \sum_{k=0}^{n} \binom{n}{k} x^k 1^{n-k} = \sum_{k=0}^{n} \binom{n}{k} x^k \tag{7}$$

Differentiating both sides of (5) with respect to x yields:

$$n(x+1)^{n-1} = \sum_{k=0}^{n} k\binom{n}{k} x^{k-1} \tag{8}$$

Plugging $x = 1$ into (6) yields:

$$n(1+1)^{n-1} = n2^{n-1} = \sum_{k=0}^{n} k\binom{n}{k} 1^{k-1} = \sum_{k=0}^{n} k\binom{n}{k} \tag{9}$$

Note: $\sum_{k=0}^{n} k\binom{n}{k} = 0 + \sum_{k=1}^{n} k\binom{n}{k} = \sum_{k=1}^{n} k\binom{n}{k}$, so we know

$$\sum_{k=1}^{n} k\binom{n}{k} = n2^{n-1} \tag{10}$$

which concludes the algebraic proof.

8. (Ross P1.31) If 8 new teachers are to be divided among 4 schools, how many divisions are possible? What if each school must receive 2 teachers?

Solution.
(a)
For the first question, there are no specifications on how many teachers must go to each school, so each teacher can go to any of the four schools. Thus, there are 4 choices for the first teacher's school, at which point there are still 4 choices for the second teacher's school, and so forth, until there are still 4 choices for the eighth teacher's school. Therefore, the total number of ways of dividing the 8 distinct teachers among the 4 distinct schools is

$$\underbrace{4 \cdot 4 \cdot 4 \cdot 4 \cdot 4 \cdot 4 \cdot 4 \cdot 4}_{8\ 4's} = 4^8 = 65,536$$

(b)
Let's number the schools $S_1, ..., S_4$. Since each school must receive 2 teachers, we must choose 2 teachers from the 8 new teachers for S_1. There are $\binom{8}{2}$ ways to do this. At this point, only 6 new teachers remain, from which we must choose 2 for S_2. There are $\binom{6}{2}$ ways to do this. Now, only 4 new teachers remain, from which we must choose 2 for S_3. There are $\binom{4}{2}$ ways to do this. Now, only 2 new teachers remain, both of which must go to S_4. There is $\binom{2}{2} = 1$ way to do this. Thus, if each school must receive 2 teachers, the the total number of ways to divide the 8 new teachers among the 4 schools is

$$\binom{8}{2}\binom{6}{2}\binom{4}{2}\binom{2}{2} = \frac{8!}{2! \cdot 6!}\frac{6!}{2! \cdot 4!}\frac{4!}{2! \cdot 2!}\frac{2!}{2! \cdot 0!} = \frac{8!}{2! \cdot 2! \cdot 2! \cdot 2! \cdot 0!} = \frac{8!}{(2!)^4} = 2,520$$

9. (Ross P1.27) Using the binomial theorem, determine the coefficient of x^6y^2 in the expansion of $(3x^2+y)^5$. Verify your answer by actually computing the expansion.

Solution.
The binomial theorem guarantees that

$$(3x^2+y)^5 = \sum_{k=0}^{5} \binom{5}{k} (3x^2)^k y^{5-k} = \sum_{k=0}^{5} \binom{5}{k} 3^k x^{2k} y^{5-k}$$

Therefore, to find the coefficient of x^6y^2 in the expansion of $(3x^2+y)^5$, we just need to find k s.t.

$$\begin{cases} 2k = 6 \\ 5-k = 2 \end{cases} \quad \text{Clearly, } 2k = 6 \implies k = 3, \text{ which also satisfies } 5-k = 2.$$

Thus, the coefficient of x^6y^2 in the expansion of $(3x^2+y)^5$ is

$$\binom{5}{k} 3^k = \binom{5}{3} 3^3 = 27 \cdot \frac{5!}{3! \cdot 2!} = 27 \cdot \frac{120}{12} = 27 * 10 = 270$$

Manually computing the expansion, we find:

$$\begin{aligned} (3x^2+y)^5 = & \quad (3x^2+y)((3x^2+y)^2)^2 = (3x^2+y)(9x^4+6x^2y+y^2)^2 \\ = & \quad (3x^2+y)(81x^8+108x^6y+54x^4y^2+12x^2y^3+y^4) \\ = & \quad 243x^{10}+405x^8y+270x^6y^2+90x^4y^3+15x^2y^4+y^5 \end{aligned}$$

which verifies that the coefficient of x^6y^2 is 270.

10. (a) The following identity is known as the Hockey Stick Identity:

$$\binom{n}{k} = \sum_{i=k}^{n} \binom{i-1}{k-1} \qquad (n \geq k).$$

Give a combinatorial argument (no computations are needed) to establish the identity.
Hint: Consider the set of numbers 1 through n. How many subsets of size k have i as their highest numbered member?

(b) Illustrate a particular case of this identity using Pascal's Triangle,

$n=0$						1					
$n=1$					1		1				
$n=2$				1		2		1			
$n=3$			1		3		3		1		
$n=4$		1		4		6		4		1	
$n=5$	1		5		10		10		5		1
$n=6$	1	6		15		20		15		6	1

While you can easily find such illustrations online, you will rob yourself of an opportunity to practice your own problem solving skills if you do so.

(c) The geometric series identity says

$$1 + y + \cdots + y^N = \frac{1 - y^{N+1}}{1 - y}.$$

Let $y = 1 + x$ and use the binomial theorem to derive a binomial coefficient identity. How does the identity relate to (a)?

Solution.
(a)
Consider the size n set $S := \{1, 2, ..., n\}$. $\binom{n}{k}$ equals the total number of size k subsets of S.

Observation 1:
For any such subset, it's highest numbered element will be some $i \in \{k, k+1, ..., n\}$.
Thus, if we calculate the number of size k subsets with i as their highest numbered member for all $i \in \{k, k+1, ..., n\}$, then add them up, this will equal the total number of size k subsets of S, $\binom{n}{k}$.

Now, let's count how many subsets of size k have a specific $i \in \{k, k+1, ..., n\}$ as their highest numbered member. Since i is the highest numbered member in the subset, all other elements must be $\in \{1, 2, ..., i-1\}$, a set with size $i-1$. Also, since the subset has size k, and i is an element of the subset, we must choose the remaining $k-1$ elements of the subset from $\{1, 2, ..., i-1\}$.
This can be done in $\binom{i-1}{k-1}$ ways. Thus, exactly $\binom{i-1}{k-1}$ subsets of S of size k have i as their highest numbered member.
Combining this with **Observation 1** and summing over all possible $i \in \{k, k+1, ..., n\}$, we obtain:

$$\binom{n}{k} = \sum_{i=k}^{n} \binom{i-1}{k-1}$$

which concludes the proof of the Hockey Stick Identity.

(b)
We will illustrate the following example:
The Hockey Stick Identity guarantees that, with $n = 6$, $k = 4$

$$\binom{6}{4} = \sum_{i=4}^{6} \binom{i-1}{3} = \binom{3}{3} + \binom{4}{3} + \binom{5}{3}$$

From Pascal's Triangle, we can quickly find that $\binom{6}{4} = 15$, $\binom{3}{3} = 1$, $\binom{4}{3} = 4$, and $\binom{5}{3} = 10$.
Now, we can easily verify that the Hockey Stick Identity holds for $n = 6$, $k = 4$:

$$\binom{6}{4} = 15 = \binom{3}{3} + \binom{4}{3} + \binom{5}{3} = 1 + 4 + 10$$

If we highlight the binomial coefficients from Pascal's Triangle that we just used for our example with $n = 6$ and $k = 4$, we find that the Hockey Stick Identity's name stems from the shape it traces through Pascal's Triangle.

n=0						1						
n=1					1		1					
n=2				1		2		1				
n=3			$\underbrace{1}_{\binom{3}{3}}$		3		3		1			
n=4		1		$\underbrace{4}_{\binom{4}{3}}$		6		4		1		
n=5	1		5		$\underbrace{10}_{\binom{5}{3}}$		10		5		1	
n=6	1	6		$\underbrace{15}_{\binom{6}{4}}$		20		15		6		1

(c)
First, multiply the right side of the geometric series identity by $1 = \frac{-1}{-1}$ to obtain:

$$1 + y + \cdots + y^N = \frac{y^{N+1} - 1}{y - 1}$$

Now, let $y = x + 1$ to obtain:

$$1 + (1 + x) + (1 + x)^2 + \cdots + (1 + x)^N = \frac{(1 + x)^{N+1} - 1}{(x + 1) - 1} = \frac{(1 + x)^{N+1} - 1}{x}$$

Note: The binomial theorem guarantees that, for all $0 \leq k \leq N$,

$$(1 + x)^i = \sum_{k=0}^{i} \binom{i}{k} x^k$$

Therefore, we know that

$$[x^k]((1 + x)^i) = \binom{i}{k}$$

While i can be any integer in $\{0, 1, ..., n\}$, for all $0 \leq i < k$, $\binom{i}{k} = 0$, so, summing over all $k \leq i \leq n$, we find that

$$[x^k](1 + (1 + x) + (1 + x)^2 + \cdots + (1 + x)^n) = \sum_{i=k}^{N} \binom{i}{k}, \ \forall\ 0 \leq k \leq N \tag{11}$$

Now, let's examine $[x^k](\frac{(1+x)^{N+1}-1}{X})$. Applying the binomial theorem, we find:

$$\begin{aligned} \frac{(1 + x)^{N+1} - 1}{X} &= \frac{(\sum_{k=0}^{N+1} \binom{N+1}{k} x^k) - 1}{x} = \frac{(\sum_{k=1}^{N+1} x^k) + \binom{N+1}{0} x^0 - 1}{x} = \frac{\sum_{k=1}^{N+1} \binom{N+1}{k} x^k}{x} \\ &= \sum_{k=1}^{N+1} \binom{N + 1}{k} x^{k-1} = \sum_{k=0}^{N} \binom{N + 1}{k + 1} x^k \end{aligned}$$

which implies that

$$[x^k](\frac{(1+x)^{N+1}-1}{X}) = \binom{N+1}{k+1} \tag{12}$$

for all $0 \leq k \leq N$. Since $1 + (1+x) + (1+x)^2 + \cdots + (1+x)^N = \frac{(1+x)^{N+1}-1}{X}$, we can combine (9) and (10) to conclude:

$$\sum_{i=k}^{N} \binom{i}{k} = \binom{N+1}{k+1} \tag{13}$$

The identity shown in (11) is directly related to the Hockey Stick Identity from part (a). To see this clearly, let $i = u - 1$ and $k = v - 1$ to find:

$$\sum_{i=k}^{N} \binom{i}{k} = \sum_{u=v}^{N+1} \binom{u-1}{v-1}$$

We can now apply the Hockey Stick Identity to verify that

$$\sum_{i=k}^{N} \binom{i}{k} = \sum_{u=v}^{N+1} \binom{u-1}{v-1} = \binom{N+1}{v} = \binom{N+1}{k+1}$$

as expected. Thus, the identity found in part (c) is just the Hockey Stick Identity from part (a), up to a change of variables.

Assignment 2

Math 407 (Swanson) – Spring 2023
Homework 1
Due Friday 1/13, 11:59pm

Name: Emerson Kahle Section: 39981

- You must upload your solutions to Gradescope as **one single, high-quality PDF**. You can convert paper-based work to a high-quality PDF using a scanning app for mobile devices, such as Adobe Scan (free, available for iOS and Android, can do multiple pages) or many others. If necessary, you can combine or merge multiple PDF's into a single PDF using a variety of services, such as Adobe Acrobat's cloud-based merge tool.
- After you upload, you must match each question with its corresponding page using Gradescope's interface. This allows graders to spend more time giving you feedback instead of hunting through submissions.
- Answers without supporting work will receive no credit. Show your work.
- You are encouraged to work together on homework, but **you must write up your solutions separately in your own words**. Copying from your fellow students or other sources is a serious academic integrity violation. In particular, you may not use "tutoring" services which simply provide answers.

- You are encouraged to typeset your solutions in LaTeX. Source code has been provided on Blackboard. Overleaf is a popular cloud-based editor.
- Problem numbers refer to the course textbook, though the problems may have been modified significantly.

1. (Ross P1.16) How many 5-card poker hands are there?

   Solution. There are 52 cards in a standard poker deck, from which we must select 5. The values and suits of the cards only, not the order in which they are selected, differentiate one poker hand from another. Thus, the total number of 5-card poker hands is:

$$\binom{52}{5} = \frac{52!}{5!47!} = \frac{52(51)(50)(49)(48)}{5(4)(3)(2)(1)} = 13 * 17 * 10 * 49 * 24 = 2,598,960$$

2. (Ross P1.28) The game of bridge is played by 4 players, each of whom is dealt 13 cards from a single standard 52-card deck. How many bridge deals are possible?

   Solution. We want to choose 4 groups of size 13 from a deck of size 52. By the definition of the multinomial coefficient, the number of ways to do this is:

$$\binom{52}{13,13,13,13} = \binom{52}{13}\binom{39}{13}\binom{26}{13}\binom{13}{13} = \frac{52!}{13!13!13!13!}$$

3. (Ross P1.10) In how many ways can 8 people be seated in a row if

   (a) there are no restrictions on the seating arrangement?

   (b) persons A and B must sit next to each other?

   (c) there are 4 men and 4 women and no 2 men or 2 women can sit next to each other?

   (d) there are 5 men and they must sit next to one another?

   (e) there are 4 married couples and each couple must sit together?

   Solution.
   (a) If there are no restrictions on the seating arrangement, then there are 8 choices for the location of the first person, 7 for the location of the second, 6 for the location of the third, 5 for the location of the fourth, 4 for the location of the fifth, 3 for the location of the sixth, 2 for the location of the seventh, and just one choice for the location of the eighth person. This results in a total of

$$8 * 7 * 6 * 5 * 4 * 3 * 2 * 1 = 8! = 40,320$$

   total possible seating arrangements.

   (b) If persons A and B must sit next to each other, then we can treat persons A and B as a combined person AB. Now, we have 7 people, and so there are 7! seating arrangements. However, 7! does not account for the position of persons A and B relative to each other. Since persons A and B account for 2 total people, there are 2! permutations of their relative location for each of the 7! possibilities, resulting in a total of

$$7! * 2! = 5,040 * 2 = 10,080$$

   total possible seating arrangements.

   (c)There are two choices for the gender of the first person. Once this choice is made, seating arrangements are only distinguished by the permutation of the group of men and the permutation of the group of women, independently. There are 4 men, so there are 4! ways to permute them. Similarly, there are 4! ways to permute the women. Since all of these choices are independent, the total number of possible seating arrangements is:

$$2 * 4! * 4! = 2 * (24)^2 = 1,152$$

   (d) Similar to part (b), let's combine the 5 men into a single person. This leaves us with 4 distinct people (the combined man and the three women), which leaves exactly 4! ways to permute them. However, 4! does not account for the positions of the men relative to each other. Since there are 5 men, for each of the 4! possibilities, there are 5! permutations of the men that result in distinct seating arrangements. Thus, the total number of possible seating arrangements is:

$$4! * 5! = 24 * 120 = 2,880$$

   (e) Similar to parts (b) and (d), we can first permute the 4 couples, which can be done in 4! ways. However, for each of these 4! possibilities, there are 2! ways to permute the first couple, 2! ways to permute the second couple, 2! ways to permute the third couple, and 2! ways to permute the fourth couple. Therefore, the total number of possible seating arrangements is:

$$4! * 2! * 2! * 2! * 2! = 4! * 2^4 = 24 * 16 = 384$$

4. A group of 11 students says their favorite animal is the cat. A separate group of 17 students says their favorite animal is the dog. Every student then flips a coin. What are the odds that 7 of the cat-loving students get Heads while 12 of the dog-loving students get Tails?

Solution.
The results of the cat-loving coin tosses are independent from the results of the dog-loving coin tosses. Therefore, we can calculate the individual probabilities of 7 cat-loving Heads and 12 dog-loving Tails, then multiply them together to arrive at a final solution. There are 2 options for each of the 11 cat-loving coin tosses, which results in a total of 2^{11} possible combinations of cat-loving coin tosses. Of these, by the definition of the binomial coefficient, there are exactly $\binom{11}{7}$ combinations of cat-loving coin tosses that result in exactly 7 Heads. Thus, the probability of exactly 7 of the cat-loving students getting Heads is:

$$\frac{\binom{11}{7}}{2^{11}} \tag{14}$$

Similarly, there are 2 options for each of the 17 dog-loving coin tosses which results in a total of 2^{17} total combinations of dog-loving coin tosses. Of these, by the definition of the binomial coefficient, there are exactly $\binom{17}{12}$ combinations of dog-loving tosses that result in exactly 12 Tails. Thus, the probability of exactly 12 dog-loving students getting Tails is:

$$\frac{\binom{17}{12}}{2^{17}} \tag{15}$$

Multiplying (1) and (2) together, we find that the total probability of 7 cat-loving students getting Heads and 12 dog-loving students getting Tails is:

$$\frac{\binom{11}{7}\binom{17}{12}}{2^{11}2^{17}} = \frac{\binom{11}{7}\binom{17}{12}}{2^{28}}$$

5. (Ross P1.29)

   (a) Expand $(x_1 + 2x_2 + 3x_3)^4$.

   (b) Interpret the coefficient of $x_1x_2x_3^2$ as the solution to some counting problem.

Solution.
(a)

$$\begin{aligned}(x_1 + 2x_2 + 3x_3)^4 &= ((x_1 + 2x_2 + 3x_3)(x_1 + 2x_2 + 3x_3))^2 \\ &= (x_1^2 + 4x_1x_2 + 6x_1x_3 + 4x_2^2 + 12x_2x_3 + 9x_3^2)^2 \\ &= x_1^4 + 8x_1^3x_2 + 12x_1^3x_3 + 24x_1^2x_2^2 + 72x_1^2x_2x_3 + 54x_1^2x_3^2 + 32x_1x_2^3 + 144x_1x_2^2x_3 \\ &\quad + 216x_1x_2x_3^2 + 108x_1x_3^3 + 16x_2^4 + 96x_2^3x_3 + 216x_2^2x_3^2 + 216x_2x_3^3 + 81x_3^4\end{aligned}$$

(b)
The coefficient of $x_1x_2x_3^2$ in $(x_1 + 2x_2 + 3x_3)^4$ is 216, which is also the solution to the following counting problem:

Suppose you roll 4 fair 3-sided dice (with values 1,2,3) and get 4 values, v_1, v_2, v_3, v_4. Let $p := v_1 * v_2 * v_3 * v_4$. Considering all possible distinct sequences of rolling 4 dice, calculate the sum of p over all sequences for which $p = 18$.

Explanation: Upon writing the prime factorization $18 = 3 * 3 * 2$, we see that the only possible combination of 4 values from $\{1, 2, 3\}$ whose product equals 18 is a combination with 1 one, 1 two, and 2 threes. Therefore, $p = 18 \implies$ the four dice rolls resulted in a total of 1 one, 1 two, and 2 threes being rolled. If we let $x_1 = 1$ is rolled, $x_2 = 2$ is rolled, and $x_3 = 3$ is rolled, then the coefficient of each term in the (unsimplified) expansion of $(x_1 + 2x_2 + 3x_3)^4$ represents an individual value of p for a specific sequence of dice rolls (v_1, v_2, v_3, v_4). Thus, each distinct sequence of dice rolls with $p = 18$ contributes the term $p = x_1 * 2x_2 * 3x_3 * 3x_3 = 18x_1x_2x_3^2$ to the expanded sum. Therefore, we can find the coefficient of $x_1x_2x_3^2$ in the simplified expansion by summing 18 once for each distinct sequence with $p = 18$. To find how many such sequences exist, we can use multinomial coefficients. We have 4 total dice rolls, from which exactly 2 must belong to the "three" group, exactly 1 must belong to the "two" group, and exactly 1 must belong to the "one" group, so we have a total of:

$$\binom{4}{2,1,1} = \frac{4!}{2!1!1!} = \frac{24}{2} = 12$$

distinct sequences of 4 dice rolls that each contribute exactly $18x_1x_2x_3^2$ to the simplified expansion of $(x_1 + 2x_2 + 3x_3)^4$. Thus, the sum of p over all distinct sequences for which $p = 18$ is:

$$\sum_{i=1}^{12} 18 = 18 * 12 = 216 = [x_1x_2x_3^2](x_1 + 2x_2 + 3x_3)^4$$

as expected.

6. (Ross TE1.13)

(a) Show that, for $n > 0$,

$$\sum_{i=0}^{n}(-1)^i\binom{n}{i} = 0.$$

Hint: Use the binomial theorem.

(b) Give a combinatorial proof of (a) when n is odd.

(c) (*Optional*) Give a combinatorial proof of (a) when n is even.

Solution.
(a) For $n > 0$, the binomial theorem guarantees that:

$$(x+1)^n = \sum_{i=0}^{n} x^i\binom{n}{i} \tag{16}$$

Plugging $x = -1$ into (3), we obtain:

$$(-1+1)^n = 0^n = 0 = \sum_{i=0}^{n}(-1)^i\binom{n}{i}$$

which concludes the proof that, for $n > 0$

$$\sum_{i=0}^{n}(-1)^i\binom{n}{i} = 0.$$

(b) Since we are summing from $i = 0$ to n, there are $n + 1$ terms in the sum. If n is odd, this means $n + 1$ is even. Therefore, we can split the sum into $\frac{n+1}{2}$ pairs of terms.
For all $0 \le i \le n$, i is either even or odd. If i is even, then $n - i$ is odd. If i is odd, then $n - i$ is even. In either case $n - i$ and i have different signs, which also means $n - i \neq i$.
This suggests, for each term corresponding to i, we should pair it with the term corresponding to $n - i$.
We can now rewrite the sum as follows:

$$\sum_{i=0}^{n}(-1)^i\binom{n}{i} = \left((-1)^0\binom{n}{0} + (-1)^n\binom{n}{n}\right) + \left((-1)^1\binom{n}{1} + (-1)^{n-1}\binom{n}{n-1}\right) + \cdots + \left((-1)^{\frac{n-1}{2}}\binom{n}{\frac{n-1}{2}} + (-1)^{\frac{n+1}{2}}\binom{n}{\frac{n+1}{2}}\right)$$

Note: By the definition of the binomial coefficient, for all $0 \le i \le n$,

$$\binom{n}{i} = \frac{n!}{i!(n-i)!} = \frac{n!}{(n-i)!(n-(n-i))!} = \binom{n}{n-i} \tag{17}$$

Since i and $n - i$ must have different signs, (4) implies that

$$(-1)^i\binom{n}{i} = -(-1)^{n-i}\binom{n}{n-i} \tag{18}$$

Now, (5) directly implies that, for all $0 \le i \le n$,

$$\left((-1)^i\binom{n}{i} + (-1)^{n-i}\binom{n}{n-i}\right) = 0 \tag{19}$$

Plugging (6) into the sum, we obtain:

$$\begin{aligned}\sum_{i=0}^{n}(-1)^i\binom{n}{i} = \quad & \left((-1)^0\binom{n}{0}+(-1)^n\binom{n}{n}\right)+\left((-1)^1\binom{n}{1}+(-1)^{n-1}\binom{n}{n-1}\right)+\cdots+ \\ & \left((-1)^{\frac{n-1}{2}}\binom{n}{\frac{n-1}{2}}+(-1)^{\frac{n+1}{2}}\binom{n}{\frac{n+1}{2}}\right) \\ = \quad & 0+0+\cdots 0 \\ = \quad & 0\end{aligned}$$

which completes the combinatorial proof for when n is odd.

(c) Since n is even, $n+1$ is odd, so we can no longer pair up each term in the sum. However, we can apply Pascal's Identity to the sum to find that:

$$\sum_{i=0}^{n}(-1)^i\binom{n}{i}=\sum_{i=0}^{n}(-1)^i\left(\binom{n-1}{i-1}+\binom{n-1}{i}\right)=\sum_{i=0}^{n}(-1)^i\binom{n-1}{i-1}+\sum_{i=0}^{n}(-1)^i\binom{n-1}{i}$$

Note: Since $\binom{n-1}{n}=0$ for all $n>0$, we know

$$\sum_{i=0}^{n}(-1)^i\binom{n-1}{i}=(-1)^n\binom{n-1}{n}+\sum_{i=0}^{n-1}(-1)^i\binom{n-1}{i}=0+\sum_{i=0}^{n-1}(-1)^i\binom{n-1}{i}=\sum_{i=0}^{n-1}(-1)^i\binom{n-1}{i}$$

Since n is even, $n-1$ is odd, so we already know from part (b) that

$$\sum_{i=0}^{n}(-1)^i\binom{n-1}{i}=\sum_{i=0}^{n-1}(-1)^i\binom{n-1}{i}=0$$

Similarly, since $\binom{n-1}{-1}=0$ for all $n>0$, we know

$$\sum_{i=0}^{n}(-1)^i\binom{n-1}{i-1}=(-1)^0\binom{n-1}{-1}+\sum_{i=1}^{n}(-1)^i\binom{n-1}{i-1}=0+\sum_{i=1}^{n}(-1)^i\binom{n-1}{i-1}=\sum_{i=1}^{n}(-1)^i\binom{n-1}{i-1}$$

In the rightmost sum, the binomial coefficients range from $\binom{n-1}{0}$ to $\binom{n-1}{n-1}$, and the exponent ranges from 1 to n. Therefore, we can rewrite the rightmost sum as:

$$\sum_{k=0}^{n-1}(-1)^{k+1}\binom{n-1}{k}$$

Now, since $n-1$ is odd, we can pair up the $(n-1)+1=n$ terms of the sum like we did in part (b). However, instead of $(-1)^i\binom{n}{i}$ and $(-1)^{n-i}\binom{n}{n-i}$, we have $(-1)^{k+1}\binom{n-1}{k}$ and $(-1)^{(n-1)-(k-1)}\binom{n-1}{(n-1)-k}$.

Note: For all $0\le k\le n-1$, $k-1$ is either even or odd. If $k-1$ is even, $(n-1)-(k-1)$ is odd. If $k-1$ is odd, $(n-1)-(k-1)$ is even. Therefore, for all $0\le k\le n-1$, $k+1$ and $(n-1)-(k-1)$ have different signs. Thus, just like in part (b), for all $0\le k\le n-1$,

$$\left((-1)^{k+1}\binom{n-1}{k}+(-1)^{(n-1)-(k-1)}\binom{n-1}{(n-1)-k}\right)=0$$

Therefore,

$$\sum_{i=0}^{n}(-1)^i\binom{n-1}{i-1}=\sum_{i=1}^{n}(-1)^i\binom{n-1}{i-1}=\sum_{k=0}^{n-1}(-1)^{k+1}\binom{n-1}{k}=0+\cdots+0=0$$

Combining all these results together, we find that:

$$\sum_{i=0}^{n}(-1)^i\binom{n}{i} = \sum_{i=0}^{n}(-1)^i\binom{n-1}{i-1} + \sum_{i=0}^{n}(-1)^i\binom{n-1}{i} = 0 + 0 = 0$$

which concludes the combinatorial proof for when n is even.

7. Recall that a permutation of $[n]$ is a word $w = w_1 \cdots w_n$ which is a rearrangement of the word $12 \cdots n$. How many permutations of [4] satisfy $w_1 > w_3$ and $w_2 > w_4$ and $w_1 > w_4$?

Solution.
Since $w_1 > w_3$ and $w_1 > w_4$, we know w_1 is greater than at least two other elements in {1,2,3,4}. This means *(i)* $w_1 \in \{3, 4\}$.
Since $w_2 > w_4$, we know w_2 is greater than at least one other element in {1,2,3,4}. This means *(ii)* $w_2 \in \{2, 3, 4\}$.
Since $w_1 > w_3$, we know that w_3 is smaller than at least one other element in {1,2,3,4}. This means *(iii)* $w_3 \in \{1, 2, 3\}$.
Since $w_1 > w_4$ and $w_2 > w_4$, we know w_4 is smaller than at least two other elements in $\{1, 2, 3, 4\}$. This means *(iv)* $w_4 \in \{1, 2\}$.
Now, we can list all 4! permutations of [4], and identify those which satisfy properties *(i), (ii), (iii),* and *(iv).*
The 24 permutations are as follows:

1 2 3 4	1 2 4 3	1 3 2 4	1 3 4 2	1 4 2 3	1 4 3 2
2 1 3 4	2 1 4 3	2 3 1 4	2 3 4 1	2 4 1 3	2 4 3 1
3 1 2 4	3 1 4 2	3 2 1 4	3 2 4 1	3 4 1 2	3 4 2 1
4 1 2 3	4 1 3 2	4 2 1 3	4 2 3 1	4 3 1 2	4 3 2 1

We can clearly see that none of the permutations from the top 2 rows satisfy property *(i)*, as $w_1 \notin \{3, 4\}$.
We can also see that none of the permutations from the bottom 2 rows of the left 2 columns satisfy property *(ii)*, as $w_2 \notin \{2, 3, 4\}$.
We can also see that none of the permutations from the third column satisfy property *(iv)*, as $w_4 \notin \{1, 2\}$.
Finally, we can see that none none of the permutations from the top three spots in column 4 satisfy property *(iii)*, as $w_3 \notin \{1, 2, 3\}$. Thus, the only permutations that satisfy all four properties are $4231, 3412, 3421, 4312, 4321$.
Clearly, all 5 of these permutations satisfy $w_1 > w_3$, $w_1 > w_4$, and $w_2 > w_4$, and we already know that the other 19 permutations do not satisfy one or more of the inequalities.
Thus, the total number of permutations of [4] that satisfy $w_1 > w_3$, $w_1 > w_4$, and $w_2 > w_4$ is 5.

8. Recall from lecture that, if $p(x) = a_0 + a_1x + \cdots + a_nx^n$, then the "reversal" of $p(x)$ is obtained by $x^n p(1/x) = a_n + a_{n-1}x + \cdots + a_0x^n$. That is, reversing the coefficients of $p(x)$ can be expressed in terms of simple algebraic operations on $p(x)$.

   (a) Let $q(x) = \sum_{k=0}^{n} ka_kx^k$. Express $q(x)$ in terms of $p(x)$ using well-known operations.

   (b) Use the fact that $1 + x + \cdots + x^n = (1 - x^{n+1})/(1 - x)$ and (a) to show that

$$1 + 2 + \cdots + n = \binom{n+1}{2}.$$

   (c) Extend (b) to prove that

$$1^3 + 2^3 + \cdots + n^3 = (1 + 2 + \cdots + n)^2.$$

Solution.
(a) First, we can rewrite $p(x)$ as a sum as follows:

$$p(x) = a_0 + a_1x + \cdots + a_nx^n = \sum_{k=0}^{n} a_kx^k$$

Differentiating both sides with respect to x, we find:

$$p'(x) = \sum_{k=0}^{n} ka_kx^{k-1}$$

Multiply both sides by x to obtain:

$$xp'(x) = \sum_{k=0}^{n} ka_kx^k = q(x)$$

Thus, we can express $q(x)$ in terms of $p(x)$ by:

$$q(x) = xp'(x) \tag{20}$$

(b) Let $a_k = 1$ for all $0 \leq k \leq n$. Then

$$p(x) = 1 + x + x^2 + ... + x^n = \frac{1 - x^{n+1}}{1 - x}$$

Differentiating both sides, we find:

$$p'(x) = 1 + 2x + 3x^2 + ... + nx^{n-1} = \frac{d}{dx}\Big(\frac{1 - x^{n+1}}{1 - x}\Big) = \frac{nx^{n+1} - (n+1)x^n + 1}{(1 - x)^2}$$

Multiplying both sides by x, we find:

$$q(x) = xp'(x) = x + 2x^2 + 3x^3 + \cdots + nx^n = \frac{nx^{n+2} - (n+1)x^{n+1} + x}{(1 - x)^2}$$

Now, let's take the limit of $q(x)$ as $x \to 1$:

$$\lim_{x\to 1} q(x) = \lim_{x\to 1} x + 2x^2 + \cdots nx^n = 1 + 2 + ... + n = \lim_{x\to 1} \frac{nx^{n+2} - (n+1)x^{n+1} + x}{(1 - x)^2}$$

Applying L'Hopital's rule, we find:

$$\lim_{x\to 1} \frac{nx^{n+2} - (n+1)x^{n+1} + x}{(1 - x)^2} = \lim_{x\to 1} \frac{n(n+2)x^{n+1} - (n+1)^2x^n + 1}{-2(1 - x)}$$

Applying L'Hopital's rule again, we find:

$$\begin{aligned}
\lim_{x\to 1}\frac{nx^{n+2}-(n+1)x^{n+1}+x}{(1-x)^2} &= \lim_{x\to 1}\frac{(n+2)(n+1)nx^n-n(n+1)^2x^{n-1}}{2}\\
&= \frac{(n+2)(n+1)(n)-n(n+1)^2}{2}\\
&= \frac{n(n^2+3n+2)-n(n^2+2n+1)}{2}\\
&= \frac{n(n+1)}{2}=\frac{n(n+1)(n-1)!}{2!(n-1)!}\\
&= \frac{(n+1)!}{2!(n-1)!}=\binom{n+1}{2}
\end{aligned}$$

Thus, we have shown that:

$$\lim_{x\to 1} q(x) = 1+2+...+n = \binom{n+1}{2} \tag{21}$$

which concludes part (b).

(c) We want to show:

$$1^3+2^3+\cdots+n^3=\sum_{k=1}^{n}k^3=(1+2+\cdots+n)^2$$

From (8), we know it is sufficient to show:

$$1^3+2^3+\cdots+n^3=\sum_{k=1}^{n}k^3=\binom{n+1}{2}^2=\left(\frac{(n+1)!}{2!(n-1)!}\right)^2=\left(\frac{(n+1)n}{2!}\right)^2=\frac{(n+1)^2n^2}{4}$$

To do this, we induct on n:
Base Case:

$$n=1 \implies \sum_{k=1}^{n}k^3=1^3=1=\frac{(1+1)^21^2}{4}=\frac{2^2*1}{4}=\frac{4}{4}=1$$

so the claim holds for the base case.

Inductive Hypothesis:
Assume that $\sum_{k=1}^{n}k^3=\frac{(n+1)^2n^2}{4}$ for all $1\le n\le j$.

Inductive Step:
Let $n=j+1$. Then

$$\sum_{k=1}^{j+1}k^3=(j+1)^3+\sum_{k=1}^{j}k^3=(j+1)^3+\frac{(j+1)^2j^2}{4}$$

by the Inductive Hypothesis. As expected,

$$\begin{aligned}
(j+1)^3+\frac{(j+1)^2j^2}{4} &= \frac{4(j^3+3j^2+3j+1)}{4}+\frac{(j+1)^2j^2}{4}=\frac{j^4+6j^3+13j^2+12j+4}{4}\\
&= \frac{(j^2+4j+4)(j^2+2j+1)}{4}=\frac{(j+2)^2(j+1)^2}{4}=\frac{((j+1)+1)^2(j+1)^2}{4}
\end{aligned}$$

The conclusion that, for all natural numbers n,

$$1^3+2^3+\cdots+n^3=\sum_{k=1}^{n}k^3=\frac{(n+1)^2n^2}{4}$$

follows by induction. This completes the proof that

$$1^3 + 2^3 + \cdots + n^3 = (1 + 2 + \cdots + n)^2 = \binom{n+1}{2}^2 = \frac{(n+1)^2 n^2}{4}$$

Assignment 3

Math 407 (Swanson) – Spring 2023
Homework 1
Due Friday 1/13, 11:59pm

Name: Emerson Kahle Section: 39981

- You must upload your solutions to Gradescope as **one single, high-quality PDF**. You can convert paper-based work to a high-quality PDF using a scanning app for mobile devices, such as Adobe Scan (free, available for iOS and Android, can do multiple pages) or many others. If necessary, you can combine or merge multiple PDF's into a single PDF using a variety of services, such as Adobe Acrobat's cloud-based merge tool.
- After you upload, you must match each question with its corresponding page using Gradescope's interface. This allows graders to spend more time giving you feedback instead of hunting through submissions.
- Answers without supporting work will receive no credit. Show your work.
- You are encouraged to work together on homework, but **you must write up your solutions separately in your own words**. Copying from your fellow students or other sources is a serious academic integrity violation. In particular, you may not use "tutoring" services which simply provide answers.
- You are encouraged to typeset your solutions in LaTeX. Source code has been provided on Blackboard. Overleaf is a popular cloud-based editor.
- Problem numbers refer to the course textbook, though the problems may have been modified significantly.

0. Assignment overview:

The "twelvefold way" consists of twelve basic counting problems. They are frequently stated in terms of the number of ways to put "balls into boxes" subject to various conditions. This assignment will explore the twelvefold way and relate it to the theory we have developed involving subsets, multisets, compositions, words, and partitions.

The balls may be distinguishable or indistinguishable; the boxes may be distinguishable or indistinguishable; and we may allow any number of balls in boxes, or each box to have at most one ball, or each box to have at least one ball. The result is $2 \cdot 2 \cdot 3 = 12$ variations on a theme.

Each of the following problems 1-7 deals with one or more of the 12 variations. As you complete these 7 problems, **write the number of the corresponding problem or sub-part in the proper square in the table below**.

	Any # balls per box	At most 1 ball/box	At least 1 ball/box
Distinguishable balls, distinguishable boxes	4: (a)	4: (b) | 3: (a), (b)	5: (a)
Distinguishable balls, indistinguishable boxes			6: (a), (b)
Indistinguishable balls, distinguishable boxes	1: (a), (b)	4: (c)	2: (a), (b)
Indistinguishable balls, indistinguishable boxes			7

1. (a) How many ways can 7 scoops of vanilla ice cream be distributed to Alice, Bob, and Stacey?
   (b) Write down an explicit general formula for distributing k scoops to n people.

Solution.
(a) There are no restrictions imposed on how many scoops Bob, Alice, and Stacey receive individually, except for the restriction that the sum of the numbers received must equal 7. Thus, the number of ways 7 indistinguishable scoops of vanilla ice cream can be distributed to Alice, Bob, and Stacey is equivalent to the number of 7 element multisubsets of a 3 element set $S = \{\text{Alice, Bob, Stacey}\}$. By the definition of the multiset coefficient, the number of ways to do this is

$$\left(\!\binom{3}{7}\!\right) = \binom{3+7-1}{7} = \binom{9}{7} = \frac{3(4)(5)(6)(7)(8)(9)}{7!} = \frac{8 \cdot 9}{2} = \frac{72}{2} = 36$$

(b) Similarly, the explicit general formula for distributing k scoops to n people is equal to the general formula for the number of k element multisubsets of an n element set. This is just the formula for the multiset coefficient:

$$\left(\!\binom{n}{k}\!\right) = \binom{n+k-1}{k} = \frac{n(n+1) \cdot (n+k-1)}{k!}$$

2. (a) How many ways can 7 scoops of vanilla ice cream be distributed to Alice, Bob, and Stacey, where each person gets at least one scoop?

   (b) Write down an explicit general formula for distributing k scoops to n people, where each person gets at least one scoop.

Solution.
(a) First, we can distribute 3 of the 7 scoops, one to each of the three people. Since the scoops are indistinguishable and each person receives the same quantity there is only one way to do this. This leaves us with 4 remaining scoops, each of which can go to any of the three people. The total number of ways to assign all 7 scoops such that each child gets at least one scoop is equal to the total number of ways to distribute the remaining 4 scoops to Alice, Bob, and Stacey, with no restrictions imposed. This is just the number of 4 element multisubsets of a 3 element set. Using the formula from 1. b), we can easily compute that there are exactly

$$\left(\binom{3}{4}\right) = \binom{3+4-1}{4} = \binom{6}{4} = \frac{3(4)(5)(6)}{4!} = \frac{5(6)}{2} = 15$$

ways 7 scoops of vanilla ice cream can be distributed to Alice, Bob, and Stacey, where each person gets at least one scoop.

(b) First, line up all k scoops next to each other, leaving $k-1$ spots in between them.

$$\underbrace{()()()...()}_{k \text{ scoops}}$$

By choosing $n-1$ of these $k-1$ spots to place dividers, we split the k scoops into n groups, where each group has at least one scoop in it. We can think of each of the n people receiving all the scoops in one group. Then, each person gets at least one scoop, and the total number of scoops adds up to k. There are exactly $\binom{k-1}{n-1}$ ways to do this. Thus, the general formula for distributing k indistinguishable scoops to n distinguishable people, where each person must receive at least one scoop, is

$$\binom{k-1}{n-1} = \frac{k(k-1)\cdots(k-n+1)}{(n-1)!}$$

Note: The formula from 2. b) is just the formula for the number of strong compositions of k indistinguishable items into n distinguishable parts. This makes sense, as the formula from 1. b) is just the formula for the number of weak compositions of k indistinguishable items into n parts, and 2. b) only changed 1. b) by imposing the restriction that all parts must have at least 1 item, which is just the distinction between strong and weak compositions.

3. (a) How many ways can 10 party guests choose from 15 possible costumes, where no two guests can choose the same costume?

   (b) Write down an explicit general formula generalizing (a).

Solution.
(a) There are 15 costume choices for the first guest, then 14 for the second guest, all the way until there are only 6 costume choices left for the 10th guest. Thus, there are a total of

$$15(14)(13)(12)(11)(10)(9)(8)(7)(6) = \frac{15!}{5!}$$

ways that 10 party guests can choose from 15 costumes such that no two guests choose the same costume.
(b) In general, if we have k party guests and n costumes, we want to compute a formula for the number of ways each of the k party guests can choose one costume such that no two guests choose the same costume. The first guest will have n choices for their costume, the second will have $n-1$, and so on, all the way until the k'th party guest has $n-k+1$ choices for their costume. Thus, the general formula for the number of ways k party guests can choose a costume from n costumes such that no two guests choose the same costume is

$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

Note: If $k > n$, then $0 \in \{n, n-1, \cdots, n-k+1\}$, so the formula outputs 0 (assuming $k, n \in \mathbb{Z}$). This makes sense, as it is impossible for k people to each choose different costumes if there are less than k costumes.

4. A *function* $f\colon A \to B$ from a set A to a set B is a rule which assigns to each input $a \in A$ some output $f(a) \in B$. For example, $f(1) = H$, $f(2) = T$, $f(3) = H$ is one particular function $f\colon \{1,2,3\} \to \{H,T\}$.

   (a) Determine the number of functions $f\colon [n] \to [x]$.

   (b) A function is *injective* if all of its outputs are distinct, i.e. for all $a_1, a_2 \in A$ with $a_1 \neq a_2$, we have $f(a_1) \neq f(a_2)$. Determine the number of injective functions $f\colon [n] \to [x]$.

   (c) A function between sets of numbers is *strictly increasing* if $a_1 < a_2$ implies $f(a_1) < f(a_2)$. Note that a strictly increasing function is necessary injective. Determine the number of strictly increasing functions $f\colon [n] \to [x]$.

Solution.
(a) We want to find the total number of functions mapping from a set of size n to a set of size x. To satisfy the definition of a function, each of the n elements from the domain must be assigned to exactly one of the x elements in the codomain. There are x choices for each of the n elements in the domain, resulting in a total of x^n distinct ways to assign each of the n elements in the domain to exactly one of the x elements in the codomain. Thus, the total number of functions $f : [n] \to [x]$ is x^n.

(b) We still need to assign each of the n elements in the domain to one of the x elements in the codomain. However, this time, assigning one element decreases the number of possible assignments for the next element by 1. We have x choices for the first element's assignment, then $x-1$ choices for the second element's assignment, and so on, until there are $x-n+1$ choices for the n'th element's assignment. All of these choices are made independently, resulting in a total of

$$x(x-1)\cdots(x-n+1)$$

ways to assign the n elements in the domain to the x elements in the codomain such that no two elements are assigned to the same place. Thus, the total number of injective functions $f : [n] \to [x]$ is $x(x-1)\cdots(x-n+1)$.

Note: If $x < n$, then $0 \in \{x, x-1, \cdots, x-n+1\}$, so there will be zero injective functions. This makes sense as you cannot assign each of the n elements from the domain to a distinct element from the codomain if the codomain has fewer total elements.

(c) We still need to assign each of the n elements in the domain to exactly one (unique) element out of the x elements in the codomain. Thus, we need to pick a subset of n elements from the codomain, to which each of its elements will be assigned exactly one element from the domain. Once we have selected this subset, we must assign the smallest number from the domain to the smallest number from the subset, then assign the second smallest number from the domain to the second smallest number from the subset, and so on, until we assign the biggest number from the domain to the biggest number in the subset. Thus, once we have selected the subset of elements from the codomain to which the elements from the domain will be assigned, there is exactly one way to actually assign them. This means the total number of ways to assign each element in the domain to an element in the codomain such that $a_1 < a_2 \implies f(a_1) < f(a_2)$ is equivalent to the total number of ways to select an n element subset of an x element set, which is $\binom{x}{n}$. Thus, the total number of increasing functions $f : [n] \to [x]$ is $\binom{x}{n}$.

Note: The result from part (c) is identical to dividing the result from part (b) by $n!$. This makes sense, as both injective and increasing functions require the n elements of the domain to be assigned to an n element subset of the codomain, but injective functions allow for all $n!$ permutations of the assignments, while increasing functions impose the restriction that elements be assigned in the (unique) strictly increasing order.

5. A function $f\colon A \to B$ is *surjective* if every $b \in B$ appears as an actual output, i.e. for all $b \in B$, there exists some $a \in A$ such that $f(a) = b$. The example function from the previous question is surjective, whereas $g\colon \{1,2,3\} \to \{H,T\}$ given by $g(1) = g(2) = g(3) = T$ is not surjective.

   (a) Determine the number of surjective functions $f\colon [4] \to [2]$.

   (b) Suppose $a_{n,k}$ is the number of surjective functions $f\colon [n] \to [k]$. Give a combinatorial proof that

$$\sum_{k=0}^{n} \binom{x}{k} a_{n,k} = x^n.$$

   Hint: let k count the number of actual outputs of an appropriate function. You may find working through the $n = 4, x = 2$ case enlightening.

Solution.
(a) It is easiest to compute the number of surjective functions $f : [4] \to [2]$ by first computing the total number of functions, then computing the number of non-surjective functions. To satisfy the definition of a function, we must assign each of the 4 elements in the domain to one of the 2 elements in the codomain. We have 2 choices for each of the 4 elements, so there are $2^4 = 16$ ways to do this. This means the total number of functions $f : [4] \to [2]$ is 16. Of these, the only ones which aren't surjective will be those that only assign elements from the domain to one of the two elements in the codomain. Suppose the domain is $D := \{a, b, c, d\}$ and the codomain is $C := \{e, f\}$. Then the only non-surjective functions $f : D \to C$ are $f_1(a) = f_1(b) = f_1(c) = f_1(d) = e$ and $f_2(a) = f_2(b) = f_2(c) = f_2(d) = f$. Thus, of the 16 possible functions, only 2 aren't surjective, so there are exactly $16 - 2 = 14$ surjective functions $f : [4] \to [2]$

(b) From 4. a), we know that the right hand side, x^n, is equivalent to the total number of functions $f : [n] \to [x]$.

Now, let's count the total number of functions $f : [n] \to [x]$ in a different way. By the definition of a function, we know that each of these functions must assign each of the n elements in the domain to one of the x elements in the codomain. However, the total number of elements in the codomain to which elements of the domain are assigned can vary. Depending on the relative sizes of n and x, the number of elements in the codomain which get assigned elements from the domain could be anything from 1 to n. For some $1 \leq k \leq n$, pick exactly k of the x elements in the codomain to get elements from the domain assigned to them. There are $\binom{x}{k}$ ways to do this. Next, count up all the functions $f : [n] \to [k]$ that assign n elements to exactly k elements. This is equal to the number of surjective functions $f : [n] \to [k]$, defined to be $a_{n,k}$. Each of these $a_{n,k}$ functions exists for each of the $\binom{x}{k}$ choices of k elements from the codomain, so there are a total of $\binom{x}{k} a_{n,k}$ functions $f : [n] \to [x]$ that assign the n elements of the domain to exactly k elements in the codomain. If we sum for all values of $1 \leq k \leq n$, we find that the total number of functions $f : [n] \to [x]$ that assign the n elements of the domain to a total of any number of elements from the codomain, which equals the total number of functions $f : [n] \to [x]$, is also equal to

$$\sum_{k=1}^{n} \binom{x}{k} a_{n,k}$$

Note: This sum is very similar to the left-hand side of the identity we want to prove, as

$$\sum_{k=0}^{n} \binom{x}{k} a_{n,k} = \binom{x}{0} a_{n,0} + \sum_{k=1}^{n} \binom{x}{k} a_{n,k} = a_{n,0} + \sum_{k=1}^{n} \binom{x}{k} a_{n,k}$$

However, $a_{n,0}$ refers to the number of surjective functions $f : [n] \to [0]$, which is 0 since there are no elements in the codomain for elements in the domain to be assigned to. Therefore,

$$\sum_{k=0}^{n} \binom{x}{k} a_{n,k} = \sum_{k=1}^{n} \binom{x}{k} a_{n,k}$$

This makes sense, as the $k = 0$ term missing from our combinatorial argument simply refers to the number of functions $f : [n] \to [x]$ that map the n elements of the domain to exactly 0 elements of the codomain, which is also 0 by the definition of a function. Thus, we have shown combinatorially that the total number of functions $f : [n] \to [x]$ is equal to:

$$\sum_{k=0}^{n} \binom{x}{k} a_{n,k} = \sum_{k=1}^{n} \binom{x}{k} a_{n,k} \ = x^n$$

which completes the proof of the identity.

6. The *Stirling numbers of the second kind*, $S(n,k)$, count the number of ways to put the integers $1, 2, \ldots, n$ into k non-empty groups, where the order of the groups does not matter. (These are called *set partitions* of $[n]$ with k non-empty *blocks*.) Unlike many of the objects we have encountered, there is no useful product formula to compute $S(n,k)$.

   (a) Compute $S(4,2)$.

   (b) Continuing the notation of the previous problem, show that

$$S(n,k) = \frac{a_{n,k}}{k!}.$$

   (c) The *falling factorial* is defined by

$$x^{\underline{n}} = x(x-1)\cdots(x-n+1).$$

   Show that the Stirling numbers of the second kind satisfy the fundamental generating function identity

$$\sum_{k=0}^{n} S(n,k)x^{\underline{k}} = x^n.$$

   Hint: You do not need to think creatively to solve this problem. You may instead combine other parts.

Solution.
(a)We need to put the numbers 1,2,3,4 into 2 nonempty groups, where the order of the groups does not matter. We can encode each distinct placement of numbers with a 4-digit binary string, where the i'th character in the string represents the group where we put the number i. For example, if we assign the numbers 1 and 2 to the same group, we could encode this as either 1100 or equivalently 0011, as the order of groups does not matter. The possible number placements can take two distinct forms:
1. One number is placed in one group and 3 numbers are placed in the other group: Since it does not matter which group has 3 numbers and which group has 1 number, we only need to choose the specific number that will stand alone in its group. We have 4 numbers, so we have 4 options, which we could encode as 1000, 0100, 0010, 0001 (or equivalently 0111, 1011, 1101, 1110)
2. Two numbers are placed in each group: Since it does not matter which group the number 1 is in, we only have to choose which number will accompany the number 1 in its group. This leaves us with 3 options, (2,3,4), which we could encode as 1100, 1010, 1001 (or equivalently 0011, 0101, 0110).
Thus, the total number of ways to put the numbers 1,2,3,4 into 2 nonempty groups, where the order of groups does not matter, is

$$S(4,2) = 3 + 4 = 7$$

encoded as 1000, 0100, 0010, 0001, 1100, 1010, 1001 (or equivalently 0111, 1011, 1101, 1110, 0011, 0101, 0110)

(b) We will give a combinatorial argument. Each set partition of [n] with k non-empty blocks can be viewed as a surjective function $f : [n] \to [k]$, as each of the k elements (blocks) in the codomain must receive at least one element (integer) from the domain. Thus, $S(n,k)$ equals the total number of distinct surjective functions $f : [n] \to [k]$ where the specific element of the codomain that a subset of the domain is assigned to *does not* matter. This is directly related to $a_{n,k}$, which counts the total number of distinct surjective functions $f : [n] \to [k]$ where the specific element of the codomain that a subset of the domain is assigned to *does* matter. Thus, $a_{n,k}$ overcounts the quantity described by $S(n,k)$ by exactly the number of permutations of the k elements of the codomain. There are exactly $k!$ ways to permute the elements of the codomain among the corresponding subsets from the domain, which completes the combinatorial argument that

$$S(n,k) = \frac{a_{n,k}}{k!}$$

(c) From 5. b), we know that

$$\sum_{k=0}^{n} \binom{x}{k} a_{n,k} = x^n$$

From 6. b), we know that

$$S(n,k) = \frac{a_{n,k}}{k!}$$

From 6. c), we know that

$$x^{\underline{k}} = x(x-1)\cdots(x-k+1)$$

Combining these three facts, we find:

$$\begin{aligned} \sum_{k=0}^{n} S(n,k) x^{\underline{k}} = & \quad \sum_{k=0}^{n} \frac{a_{n,k}}{k!} x(x-1)\cdots(x-k+1) = \sum_{k=0}^{n} a_{n,k} \frac{x(x-1)\cdots(x-k+1)}{k!} \\ = & \quad \sum_{k=0}^{n} a_{n,k} \frac{x!}{k!(x-k)!} = \sum_{k=0}^{n} a_{n,k} \binom{x}{k} = x^n \end{aligned}$$

which completes the proof that the Stirling numbers of the second kind satisfy the fundamental generating function identity

$$\sum_{k=0}^{n} S(n,k) x^{\underline{k}} = x^n.$$

7. How many ways can you create words using the letters U, S, C where

   (i) each letter is used at least once;

   (ii) the total length is 6;

   (iii) at least as many U's are used as S's;

   (iv) at least as many S's are used as C's;

   (v) and the word is lexicographically last among all of its rearrangements.

Solution.
Let $n_U = \#$ of U's in the word, $n_S = \#$ of S's in the word, and $n_C = \#$ of C's in the word. Based on the listed requirements, we know that

$$\begin{cases} n_U + n_S + n_C = 6 \\ n_U \geq n_S \geq n_C \\ n_U, n_S, n_C \geq 1 \end{cases}$$

This leaves us with three distinct possibilities for (n_U, n_S, n_C), which are as follows:

$$\begin{cases} (2,2,2) \\ (3,2,1) \\ (4,1,1) \end{cases}$$

Regardless of n_U, n_S, n_C, since each word is lexicographically last among all of its rearrangements, it must have all of its U's before all of its S's and all of its S's before all of its C's.
For $(n_U, n_S, n_C) = (2,2,2)$, this leaves UUSSCC as the only possibility.
For $(n_U, n_S, n_C) = (3,2,1)$, this leaves UUUSSC as the only possibility.
For $(n_U, n_S, n_C) = (4,1,1)$, this leaves UUUUSC as the only possibility.

Thus, the only possible words are UUSSCC, UUUSSC, and UUUUSC, so you can create a total of 3 words using the letters U, S, C under the 5 requirements.

8. (Not part of the twelvefold way.) Suppose you roll 10 4-sided dice.

   (a) What are the odds the sum is 10?

   (b) What are the odds the sum is 18?

   (c) What is the most likely sum?

Solution.
(a) For each of the 10 rolls, there are 4 different possible values that can be rolled, so there are 4^{10} total sequences of dice rolls.
Of these, only 1 sequence results in a sum of 10: (1,1,1,1,1,1,1,1,1,1)
Thus, the probability that the sum of 10 4-sided dice rolls is 10 is

$$\frac{1}{4^{10}} \approx 0.000095\%$$

(b) First, we must determine all possible combinations of 10 4-sided dice rolls which result in a sum of 18. Then, for each such combination, we can use multinomial coefficients to determine how many distinct sequences result from that combination. We can sum this quantity for each such combination to determine the total number of dice roll sequences that result in a sum of 18.

Let $k_1 = \#$ of 1's rolled, $k_2 = \#$ of 2's rolled, $k_3 = \#$ of 3's rolled, and $k_4 = \#$ of 4's rolled. Then we can will denote a single combination of 10 rolls by (k_1, k_2, k_3, k_4). The total number of sequences resulting in any such combination is $\binom{10}{k_1,k_2,k_3,k_4} = \frac{10!}{k_1!k_2!k_3!k_4!}$.
The combinations with sums of 18 are as follows:

$$\text{(Zero 4's)} \begin{cases} (3,6,1,0) \\ (4,4,2,0) \\ (5,2,3,0) \\ (6,0,4,0) \\ (2,8,0,0) \end{cases} \quad \text{(One 4)} \begin{cases} (4,5,0,1) \\ (5,3,1,1) \\ (6,1,2,1) \end{cases} \quad \text{(Two 4's)} \begin{cases} (6,2,0,2) \\ (7,0,1,2) \end{cases}$$

The corresponding multinomial coefficients are as follows:

$$\text{(Zero 4's)} \begin{cases} \binom{10}{3,6,1,0} = \frac{10!}{3!6!} \\ \binom{10}{4,4,2,0} = \frac{10!}{4!4!2!} \\ \binom{10}{5,2,3,0} = \frac{10!}{5!2!3!} \\ \binom{10}{6,0,4,0} = \frac{10!}{6!4!} \\ \binom{10}{2,8,0,0} = \frac{10!}{2!8!} \end{cases} \quad \text{(One 4)} \begin{cases} \binom{10}{4,5,0,1} = \frac{10!}{4!5!} \\ \binom{10}{5,3,1,1} = \frac{10!}{5!3!} \\ \binom{10}{6,1,2,1} = \frac{10!}{6!2!} \end{cases} \quad \text{(Two 4's)} \begin{cases} \binom{10}{6,2,0,2} = \frac{10!}{6!2!2!} \\ \binom{10}{7,0,1,2} = \frac{10!}{7!2!} \end{cases}$$

Thus, the total number of 10 roll sequences whose rolls sum to 18 is

$$\frac{10!}{3!6!} + \frac{10!}{4!4!2!} + \frac{10!}{5!2!3!} + \frac{10!}{6!4!} + \frac{10!}{2!8!} + \frac{10!}{4!5!} + \frac{10!}{5!3!} + \frac{10!}{6!2!} + \frac{10!}{6!2!2!} + \frac{10!}{7!2!} = 17,205$$

Therefore, the probability that a sequence of 10 4-sided dice rolls results in a sum of 18 is

$$\frac{17,205}{4^{10}} \approx 1.64\%$$

Note: I found the various combinations with sums of 18 by solving the system

$$\begin{cases} k_1 + k_2 + k_3 + k_4 = 10 \\ k_1 + 2k_2 + 3k_3 + 4k_4 = 18 \end{cases} \quad \text{under the restriction } k_1, k_2, k_3, k_4 \geq 0.$$

(c) Let $X =$ an individual roll of a 4-sided die. Then there is a $\frac{1}{4}$ probability of rolling each of the 4 values, 1,2,3,4. Thus,

$$E(X) = \frac{1}{4} + \frac{2}{4} + \frac{3}{4} + \frac{4}{4} = \frac{10}{4} = 2.5$$

Thus, if we let X_{10} = rolling 10 4-sided dice, then, since the rolls are independent,

$$E(X_{10}) = 10 * E(X) = 10 * 2.5 = 25$$

Since 25 is an integer between 10 and 40, it is a valid sum that can result from a sequence of 10 4-sided dice rolls. Moreover, since 25 is the expected sum from 10 4-sided dice rolls, it is also the most likely sum from the 10 rolls. This makes sense, as 10 is the lowest possible sum of 10 rolls, 40 is the highest possible sum of 10 rolls, and 25 is in the exact center of $[10, 40]$. Thus, the most likely sum from 10 4-sided dice rolls is 25. This could also be shown by expanding $(x + x^2 + x^3 + x^4)^{10}$, whose largest coefficient would be on the term x^{25}.

9. (Not part of the twelvefold way.) How many solutions are there to

$$x + y + z = 10$$

where x, y, z are integers satisfying $x \geq -3, y \geq 0, z \geq 3$?
Solution.
We can manipulate the equation to make counting more familiar. Note that, since $3 - 3 = 0$,

$$x + y + z = x + y + z + 3 - 3 = (x + 3) + y + (z - 3) = 10$$

Since $x \geq -3$, we know $x + 3 \geq 0$. Similarly, since $z \geq 3$, we know $z - 3 \geq 0$. If we let $x_1 = x + 3$ and $z_1 = z - 3$, then we can rewrite the initial equation as

$$x_1 + y + z_1 = 10$$

with the restriction $x_1, y, z_1 \geq 0$. At this point, we can think of the problem as the weak combinations of 10 into 3 parts, which is

$$\binom{n + k - 1}{k - 1} = \binom{12}{2} = \frac{12!}{10!2!} = \frac{132}{2} = 66$$

Thus, there are exactly 66 solutions to

$$x_1 + y + z_1$$

where $x_1, y, z_1 \geq 0$ are integers. Since

$$x + y + z = 10, \text{ where } x \geq -3, y \geq 0, z \geq 3$$

is equivalent to

$$x_1 + y + z_1 = 10, \text{ where } x_1, y, z_1 \geq 0$$

the two equations have the same number of solutions.
Thus, there are also exactly 66 solutions (x, y, z) to

$$x + y + z = 10$$

where x, y, z are integers satisfying $x \geq -3, y \geq 0, z \geq 3$.

Assignment 4

Math 407 (Swanson) – Spring 2023
Homework 1
Due Friday 1/13, 11:59pm

Name: Emerson Kahle Section: 39981

- You must upload your solutions to Gradescope as **one single, high-quality PDF**. You can convert paper-based work to a high-quality PDF using a scanning app for mobile devices, such as Adobe Scan (free, available for iOS and Android, can do multiple pages) or many others. If necessary, you can combine or merge multiple PDF's into a single PDF using a variety of services, such as Adobe Acrobat's cloud-based merge tool.

- After you upload, you must match each question with its corresponding page using Gradescope's interface. This allows graders to spend more time giving you feedback instead of hunting through submissions.
- Answers without supporting work will receive no credit. Show your work.
- You are encouraged to work together on homework, but **you must write up your solutions separately in your own words**. Copying from your fellow students or other sources is a serious academic integrity violation. In particular, you may not use "tutoring" services which simply provide answers.
- You are encouraged to typeset your solutions in LaTeX. Source code has been provided on Blackboard. Overleaf is a popular cloud-based editor.
- Problem numbers refer to the course textbook, though the problems may have been modified significantly.

1. Let c_n denote the number of partitions of n into *distinct* parts. For example, $\lambda = (3,3,2,1)$ has 3 repeated twice, so it should not be counted in c_9, whereas $\lambda = (5,3,2,1)$ does contribute to c_{11}.

   (a) Show that
   $$\sum_{n=0}^{\infty} c_n x^n = \prod_{i=1}^{\infty}(1+x^i).$$
   (*Hint:* adapt the proof of Euler's product formula from lecture.)

   (b) Argue that
   $$c_n = [x^n]\prod_{i=1}^{n}(1+x^i).$$

   (c) Use the formula in the previous part to compute c_5.

Solution.
(a) A partition of n is a partition into distinct parts if and only if it satisfies the following two conditions, where e_i = the # of size i parts in the partition, for all natural numbers i:
(i) $0 \le e_1, e_2, \cdots \le 1$
(ii) $e_1 + 2e_2 + \cdots = n$
Therefore, we can rewrite c_n as a summation as follows:

$$c_n = \sum_{\substack{0\le e_1,e_2,\cdots\le 1\\ e_1+2e_2+\cdots=n}} 1$$

Now, we can rewrite the infinite sum from (a) as:

$$\sum_{n=0}^{\infty} c_n x^n = \sum_{n=0}^{\infty}\Big(\sum_{\substack{0\le e_1,e_2,\cdots\le 1\\ e_1+2e_2+\cdots=n}} 1\Big)x^n$$

Since $e_1 + 2e_2 + \cdots = n$ in every term in the innermost sum, and x does not depend on $e_1, e_2, \cdots$ we can move x^n inside the innermost sum as follows:

$$\sum_{n=0}^{\infty} c_n x^n = \sum_{n=0}^{\infty}\Big(\sum_{\substack{0\le e_1,e_2,\cdots\le 1\\ e_1+2e_2+\cdots=n}} 1\Big)x^n = \sum_{n=0}^{\infty}\sum_{\substack{0\le e_1,e_2,\cdots\le 1\\ e_1+2e_2+\cdots=n}} x^{e_1+2e_2+\cdots}$$

Note: The innermost sum counts over all combinations of $e_1, e_2, \cdots$ such that $0 \le e_1, e_2, \cdots \le 1$ and $e_1 + 2e_2 + \cdots = n$, and the outermost sum counts one of these innermost sums for all non-negative integers n. Thus, we can remove the need for the outermost sum by simply removing the restriction $e_1 + 2e_2 + \cdots = n$. This allows for all combinations of $0 \le e_1, e_2, \cdots \le 1$ such that $e_1 + 2e_2 + \cdots = n$ for any non-negative integer n to be counted with just one sum. Thus, we know that:

$$\sum_{n=0}^{\infty}\sum_{\substack{0\le e_1,e_2,\cdots\le 1\\ e_1+2e_2+\cdots=n}} x^{e_1+2e_2+\cdots} = \sum_{0\le e_1,e_2,\cdots\le 1} x^{e_1+2e_2+\cdots}$$

Since $x^{e_1+2e_2+\cdots} = (x^{e_1})(x^{2e_2})\cdots$, we know that:

$$\sum_{0\le e_1,e_2,\cdots\le 1} x^{e_1+2e_2+\cdots} = \sum_{0\le e_1,e_2,\cdots\le 1}(x^{e_1})(x^{2e_2})\cdots = \Big(\sum_{e_1=0}^{1} x^{e_1}\Big)\Big(\sum_{e_2=0}^{1} x^{2e_2}\Big)\cdots$$

Now, we can easily compute each sum individually to find that:

$$\Big(\sum_{e_1=0}^{1} x^{e_1}\Big)\Big(\sum_{e_2=0}^{1} x^{2e_2}\Big)\cdots = (1+x^1)(1+x^2)\cdots = \prod_{i=1}^{\infty}(1+x^i)$$

Thus, we have shown that:

$$\sum_{n=0}^{\infty} c_n x^n = \sum_{0 \le e_1, e_2, \cdots \le 1} x^{e_1 + 2e_2 + \cdots} = \Big(\sum_{e_1=0}^{1} x^{e_1}\Big)\Big(\sum_{e_2=0}^{1} x^{2e_2}\Big) \cdots = \prod_{i=1}^{\infty}(1 + x^i)$$

which completes the proof that

$$\sum_{n=0}^{\infty} c_n x^n = \prod_{i=1}^{\infty}(1 + x^i).$$

(b) We want to show that

$$c_n = [x^n] \prod_{i=1}^{n}(1 + x^i)$$

We can clearly see that

$$c_n = [x^n] \sum_{n=0}^{\infty} c_n x^n$$

From part (a), since $\sum_{n=0}^{\infty} c_n x^n = \prod_{i=1}^{\infty}(1 + x^i)$, this implies that

$$c_n = [x^n] \prod_{i=1}^{\infty}(1 + x^i)$$

Thus, it suffices to show that, for all natural numbers n,

$$[x^n] \prod_{i=1}^{\infty}(1 + x^i) = [x^n] \prod_{i=1}^{n}(1 + x^i)$$

Let $b_i = (1 + x^i)$ for all natural numbers i.
Consider the expansion of $\prod_{i=1}^{\infty} b_i$, which is a sum of infinitely many terms.
c_n is equal to the number of such terms that equal x^n.
Each of the terms in the expanded infinite sum has exactly one factor from each b_i in the initial infinite product.
For all b_i, the factor that b_i contributes to a given term in the expanded sum is either 1 or x^i.
Thus, for all b_i, the factor from b_i can only increase or not change the power of x in a specific term of the expanded infinite sum.
It follows that, if some b_j s.t. $j > n$ contributes x^j to a specific term in the expanded infinite sum, then that term can never equal x^n.
Thus, each b_j s.t. $j > n$ must contribute a 1 as its factor to each term in the expanded infinite sum that equals x^n.
Thus, for each such term in the expanded infinite sum, the only b_i's that possibly contribute *meaningful* factors (i.e. factors that aren't 1) to that term are b_i s.t. $1 \le i \le n$.
Thus, for each term that equals x^n in the expansion of $\prod_{i=1}^{\infty} b_i$, there is exactly one term that equals x^n in the expansion of $\prod_{i=1}^{n} b_i$.
In fact, the corresponding terms each received the same factors from all b_i s.t $1 \le i \le n$, but the terms from $\prod_{i=1}^{n} b_i$ are just missing the infinitely many 1's contributed by the infinitely many b_i s.t. $i > n$.
Thus, the number of copies of x^n in $\prod_{i=1}^{\infty} b_i$ is exactly the same as the number of copies of x^n in $\prod_{i=1}^{n} b_i$.
Thus, we have shown that

$$[x^n] \prod_{i=1}^{\infty}(1 + x^i) = [x^n] \prod_{i=1}^{n}(1 + x^i)$$

so we know that

$$c_n = [x^n] \sum_{n=0}^{\infty} c_n x^n = [x^n] \prod_{i=1}^{\infty}(1 + x^i) = [x^n] \prod_{i=1}^{n}(1 + x^i)$$

which completes the argument that

$$c_n = [x^n] \prod_{i=1}^{n} (1 + x^i)$$

(c) From part (b), we know that

$$c_5 = [x^5] \prod_{i=1}^{5} (1 + x^i) = [x^5] \big((1 + x)(1 + x^2)(1 + x^3)(1 + x^4)(1 + x^5) \big)$$

Now, we can expand this product to directly calculate c_5.

$$\begin{aligned}
& (1 + x)(1 + x^2)(1 + x^3)(1 + x^4)(1 + x^5) \\
= \; & (1 + x + x^2 + x^3)(1 + x^3)(1 + x^4)(1 + x^5) \\
= \; & (1 + x + x^2 + 2x^3 + x^4 + x^5 + x^6)(1 + x^4)(1 + x^5) \\
= \; & (1 + x + x^2 + 2x^3 + 2x^4 + 2x^5 + 2x^6 + 2x^7 + x^8 + x^9 + x^{10})(1 + x^5) \\
= \; & (1 + x + x^2 + 2x^3 + 2x^4 + 3x^5 + 3x^6 + 3x^7 + 3x^8 + 3x^9 + 3x^{10} + 2x^{11} + 2x^{12} + x^{13} + x^{14} + x^{15})
\end{aligned}$$

Now, we can clearly see that

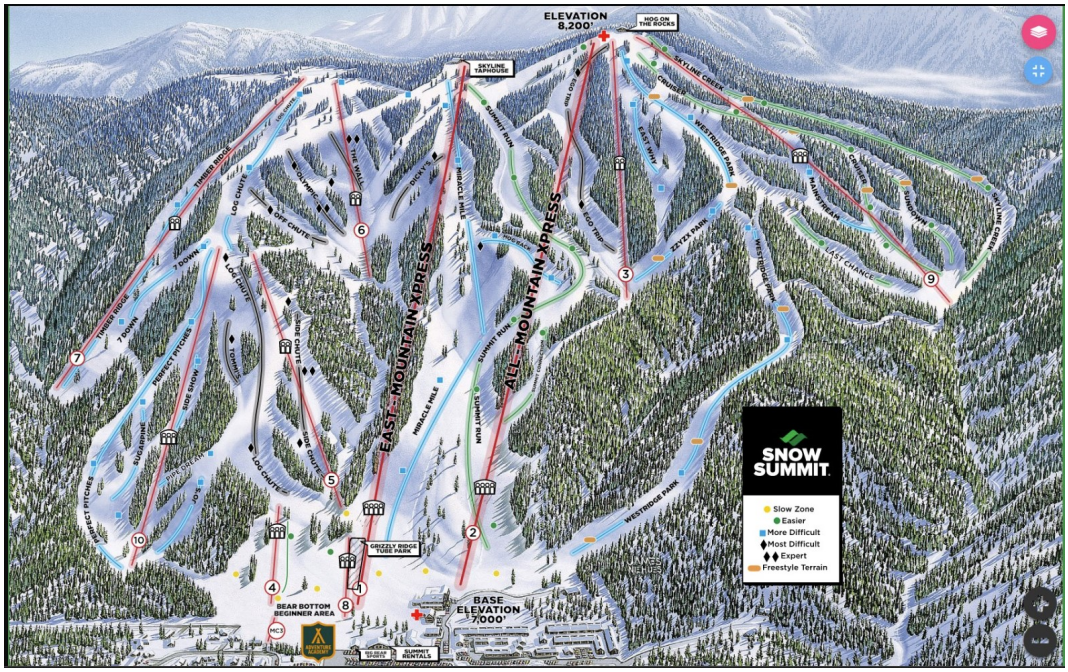
$$c_5 = [x^5] \prod_{i=1}^{5} (1 + x^i) = 3$$

We can verify that this result is correct by listing all possible partitions of 5 into distinct parts, which are: (5), (4,1), and (3,2). As expected, there are exactly 3 such partitions.

2. You are on a chair lift at a certain Southern California ski resort. You notice a message on one of the support poles:

The message says, "How many different ways down can you find using chairs 5, 6, 7, & 10 in combination ?"

You quickly look at the resort's trail map:



What is your answer to the message?

Solution.
First, we must define a few terms and state a few assumptions in order to arrive at a conclusive quantifiable solution.
Assumption 1: Any 'way down' must start at one of the mountain's peaks and end at the base of the mountain
Assumption 2: For a 'way down' to 'use a combination of chairs 5, 6, 7, and 10,' these chairs must provide the most direct path to/from the start/end of the 'way down.'

Let's call any sequence of trails a *route*. Let's call any 'way down using chairs 5,6,7, & 10 in combination' a *valid route*. We will say that a *route* is a *valid route* if and only if it satisfies the following conditions:

(1) The first trail in the *route* starts at either the top of chair 6 or 7.

(2) The last trail in the *route* ends at the bottom of either chair 5 or 10.

(3) Every trail in the *route* (besides the last trail) connects directly to the next trail in the *route*.

(4) No trail in the route starts or ends at a chair other than chair 5, 6, 7, or 10.

Now, we will provide a rationale for each rule:

(1) The tops of both chairs 6 and 7 end at one of the mountain's three peaks. By imposing this restriction, we ensure that all *valid routes* start at one of the mountain's peaks, as required by *Assumption 1*.

(2) The bottoms of both chairs 5 and 10 are at the base of the mountain. By imposing this restriction, we ensure that all *valid routes* end at the mountain's base, as required by *Assumption 1*.

(3) For our *route* to be a valid 'way down' the mountain, it must consist of connected trails so that one could ski down our *route* in the specified order without needing to take any other trails. This restriction ensures that there are no discontinuities in any *valid route*.

(4) If a trail starts or ends at a chair other than 5, 6, 7, or 10, then it would be easiest to ski to/from that trail using a chair other than 5, 6, 7, or 10. However, we want all *valid routes* to use a 'combination of chairs 5, 6, 7, or 10.' By imposing this restriction, we ensure that chairs 5, 6, 7, and 10 combine to provide the most direct path to/from the start/end of every *valid route*, as required by *Assumption 2*.

Now that we have identified assumptions, defined rules, and provided rationales, we can count the number of *valid routes* under these conditions.
First, we can note that condition (1) forces the first trail in any *valid route* to be the wall, olympic, log chute, or timber ridge.
Also, we can note that condition (2) forces the last trail in any *valid route* to be side chute, log chute, jo's, pipe dream, or perfect pitches.
Now, for each potential starting trail, let's list the potential *valid routes*.
i) We start at the wall:
<the wall, side chute>
so only 1 *valid route* starts at the wall.
ii) We start at olympic:
<olympic, side chute>
so only 1 *valid route* starts at olympic.
iii) We start at log chute:
<log chute>, <log chute, olympic, side chute>, <log chute, off chute, side chute>,
<log chute, side chute>, <log chute, jo's>, <log chute, pipe dream>,
<log chute, tommi's, log chute>, <log chute, tommi's, log chute, side chute>,
<log chute, tommi's, jo's>, <log chute, tommi's, pipe dream>,
<log chute, side show, pipe dream>, <log chute, side show, sugarpine, perfect pitches>, <log chute,

perfect pitches>, <log chute, perfect pitches, side show, pipe dream>,
<log chute, perfect pitches, side show, sugarpine, perfect pitches>,
<log chute, perfect pitches, sugarpine, perfect pitches>,
<log chute, 7 down, perfect pitches>, <log chute, 7 down, sugarpine, perfect pitches>
so exactly 18 *valid routes* start at log chute.
iv) We start at timber ridge:
<timber ridge, 7 down, perfect pitches>,
<timber ridge, 7 down, sugarpine, perfect pitches>
so exactly 2 *valid routes* start at timber ridge.
Adding the number of *valid routes* for each of the possible starting trails, we find that there are exactly:

$$1 + 1 + 18 + 2 = 22$$

total *valid routes*.
Thus, the total number of 'ways down using chairs 5,6,7, & 10 in combination' is 22.

3. (Ross P2.1) A box contains 3 marbles: 1 red, 1 green, and 1 blue. Consider an experiment that consists of taking 1 marble from the box and then replacing it in the box and drawing a second marble from the box. Describe the sample space. Repeat when the second marble is drawn without replacing the first marble.

Solution.
(a) We want to describe the sample space for the experiment when the first marble is replaced. For the first marble chosen, there are 3 possible choices, red, green, and blue. Since the first marble is replaced into the box, there are also the same 3 possible choices for the second marble. Thus, the sample space is:
S = {red, green, blue}^{2} = {(red,red), (red, green), (red, blue), (green, red), (green, green), (green, blue), (blue, red), (blue, green), (blue, blue)}
Here, the size of the sample space is $|S| = |\{\text{red, green, blue}\}|^2 = 3^2 = 9$.

(b) We want to describe the sample space for the experiment when the first marble is not replaced. There are 3 possible choices for the first marble, red, green, and blue. This leaves the two colors that were not chosen as the two choices for the color of the second marble. Thus, the sample space is:
S' = {(red, green), (red, blue), (green, red), (green, blue), (blue, red), (blue, green)}
Here, the size of the sample space is $|S'| = 6$.
Note: $S' = S - \{(\text{red, red}), (\text{green, green}), (\text{blue, blue})\}$, as the only events from S that disallowing replacement prevents are those in which the first and second marbles share the same color.

4. In an experiment, a die is rolled repeatedly until a 6 appears, at which point the experiment stops. What is the sample space of this experiment? Let E_n denote the event that n rolls are necessary to complete the experiment. What points of the sample space are contained in E_n? What is $(\cup_{n=1}^{\infty} E_n)^c$?

Solution.
Note: For the duration of this problem $\{1,2,3,4,5\}^i$ denotes the set of all possible sequences of i elements of $\{1,2,3,4,5\}$.
Similarly, $\{\{1,2,3,4,5\}^i 6\}$ denotes the set of all possible sequences of i elements of $\{1,2,3,4,5\}$, followed by one 6.

(a) We want to find the sample space of this experiment.
Note: The sample space contains all possible outcomes of the experiment, and a given outcome can take anywhere from $1 \to \infty$ dice rolls to be completed.
Let $f(6) : \{6\} \to \{1,2,\cdots\}$, where $f(6) =$ the roll on which 6 first appears, with $f(6) = \infty$ meaning that 6 never appears during the experiment.
We can identify the set of possible outcomes for each value of $f(6)$, and the union of these sets will be our sample space.

For all $i \in \mathbb{N}$, if $f(6) = i$, then we know the i'th roll was a 6, so the experiment only takes i dice rolls. We also know that 6 did not appear in the first $i-1$ dice rolls, so the set of possible outcomes for the first $i-1$ dices rolls is

$$\{1,2,3,4,5\}^{i-1}$$

Thus, the set of possible outcomes when $f(6) = i$ is

$$\{\{1,2,3,4,5\}^{i-1}6\}$$

for all $i \in \mathbb{N}$.
For example, if $f(6) = 1$, then the set of possible outcomes is

$$\{\{1,2,3,4,5\}^{1-1}6\} = \{6\}$$

If $f(6) = 2$, then the set of possible outcomes of the experiment is:

$$\{\{1,2,3,4,5\}^{2-1}6\} = \{\{1,2,3,4,5\}6\}$$

If $f(6) = 3$, then the set of possible outcomes of the experiment is:

$$\{\{1,2,3,4,5\}^{3-1}6\} = \{\{1,2,3,4,5\}^2 6\}$$

This is true for all $i \in \mathbb{N}$, but we also need to consider what happens when $i \to \infty$.
This represents all outcomes where we never roll a 6, which is equivalent to all infinite sequences of elements from $\{1,2,3,4,5\}$.
Thus, if $f(6) = \infty$, then the set of possible outcomes of the experiment is:

$$\{1,2,3,4,5\}^{\infty}$$

Now that we have identified the set of possible outcomes for all possible values of $f(6)$, we can take their infinite union to find the sample space:

$$\begin{aligned} S = & \quad \Big(\{6\} \cup \{\{1,2,3,4,5\}6\} \cup \{\{1,2,3,4,5\}^2 6\} \cup ...\Big) \cup \{1,2,3,4,5\}^{\infty} \\ = & \quad \Big(\bigcup_{i=0}^{\infty} \{\{1,2,3,4,5\}^i 6\}\Big) \cup \{1,2,3,4,5\}^{\infty} \end{aligned}$$

(b) We want to describe the contents of E_n for all n.
Note: E_n is nonempty for all $i \in \mathbb{N}$, as there are there are outcomes of the experiment for which i rolls

are necessary for all $i \in \mathbb{N}$.
Also, for an outcome $e \in E_n$, the experiment must stop after n dice rolls, so a 6 must be rolled for the first time on the n'th dice roll. Therefore, all outcomes in which a 6 is never rolled (represented in part (a) by $f(6) = \infty$), are not included in any event E_n.

Just like when $f(6) = n$ for some $n \in \mathbb{N}$, for an outcome $e \in E_n$, the first $n-1$ rolls must be some sequence of elements from $\{1,2,3,4,5\}$, and the n'th roll must be a 6. Thus, for all $n \in \mathbb{N}$ (the only values for which $E_n \neq \emptyset$), we know that

$$E_n = \{\{1,2,3,4,5\}^{n-1}6\}$$

And for all $n \notin \mathbb{N}$,

$$E_n = \emptyset$$

(c) We want to find $(\cup_{n=1}^{\infty} E_n)^c$.
Applying the formula from part (b), we find:

$$\bigcup_{n=1}^{\infty} E_n = \bigcup_{n=1}^{\infty} \{\{1,2,3,4,5\}^{n-1}6\} = \bigcup_{n=0}^{\infty} \{\{1,2,3,4,5\}^{n}6\}$$

From part (a), we know that $S = \left(\bigcup_{n=0}^{\infty} \{\{1,2,3,4,5\}^n 6\}\right) \cup \{1,2,3,4,5\}^{\infty}$.
Thus, we can clearly see that

$$\begin{aligned} \bigcup_{n=1}^{\infty} E_n = & \quad \bigcup_{n=0}^{\infty} \{\{1,2,3,4,5\}^n 6\} = \left(\left(\bigcup_{n=0}^{\infty} \{\{1,2,3,4,5\}^n 6\}\right) \cup \{1,2,3,4,5\}^{\infty}\right) - \{1,2,3,4,5\}^{\infty} \\ = & \quad S - \{1,2,3,4,5\}^{\infty} \end{aligned}$$

Since our universe is just the sample space, S, $\left(\bigcup_{n=1}^{\infty} E_n\right)^c = S - \bigcup_{n=1}^{\infty} E_n$. Therefore,

$$\begin{aligned} \left(\bigcup_{n=1}^{\infty} E_n\right)^c &= S - \bigcup_{n=1}^{\infty} E_n = S - (S - \{1,2,3,4,5\}^{\infty}) = S - (S\,(\{1,2,3,4,5\}^{\infty})^c) \\ &= S(S\,(\{1,2,3,4,5\}^{\infty})^c)^c = S(S^c \cup (\{1,2,3,4,5\}^{\infty})^{cc}) = SS^c \cup S\,(\{1,2,3,4,5\}^{\infty})^{cc} \\ &= \emptyset \cup S\{1,2,3,4,5\}^{\infty} = \{1,2,3,4,5\}^{\infty} \end{aligned}$$

Thus, we have found that

$$\left(\bigcup_{n=1}^{\infty} E_n\right)^c = \{1,2,3,4,5\}^{\infty}$$

5. (Ross P2.3)

(a) Two dice are thrown. Let E be the event that the sum of the dice is odd, let F be the event that at least one of the dice lands on 1, and let G be the event that the sum is 5. Describe the events $EF, E \cup F, FG, EF^c$, and EFG. (Recall that $EF = E \cap F$.)

(b) Assuming the dice are fair, what are the probabilities of each of the above events?

Solution.
(a) First, we should explicitly describe the events E, F, and G themselves:
E consists of all sequences of dice rolls that include 1 odd and 1 even:

$$E = \{(1,2),(1,4),(1,6),(2,1),(2,3),(2,5),(3,2),(3,4),(3,6),\\ (4,1),(4,3),(4,5),(5,2),(5,4),(5,6),(6,1),(6,3),(6,5)\}$$

F consists of all sequences of dice rolls that include at least one 1.

$$F = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(3,1),(4,1),(5,1),(6,1)\}$$

G consists of all sequences of dice rolls that result in a sum of 5.

$$G = \{(1,4),(2,3),(3,2),(4,1)\}$$

Now, we can describe the events $EF, E \cup F, FG, EF^c$, and EFG one by one using the definitions of the various set operations.
i) $EF = E \cap F =$ the set of all sequences of dice rolls (r_1, r_2) s.t. $(r_1, r_2) \in E$ and $(r_1, r_2) \in F$. In words, this is equivalent to the set of all sequences of dice rolls that both include at least one 1 and have an odd sum of dice. The set of such sequences (r_2, r_2) is:

$$EF = \{(1,2),(1,4),(1,6),(2,1),(4,1),(6,1)$$

ii) $E \cup F =$ the set of all sequences of dice rolls (r_1, r_2) s.t. $(r_1, r_2) \in E$ or $(r_1, r_2) \in F$, or both. In words, this is equivalent to the set of all sequences of dice rolls that either include at least one 1 or have an odd sum, or both. The set of all such sequences (r_1, r_2) is:

$$E \cup F = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,3),(2,5),(3,1),(3,2),(3,4),\\ (3,6),(4,1),(4,3),(4,5),(5,1),(5,2),(5,4),(5,6),(6,1),(6,3),(6,5)\}$$

iii) $FG = F \cap G =$ the set of all sequences of dice rolls (r_1, r_2) s.t. $(r_1, r_2) \in F$ and $(r_1, r_2) \in G$. In words, this is equivalent to the set of all sequences of dice rolls that both include at least one 1 and whose sum of dice is 5. The set of all such sequences (r_1, r_2) is:

$$FG = \{(1,4),(4,1)\}$$

iv) $EF^c = E \cap F^c = E \cap (S - F) =$ the set of all sequences of dice rolls (r_1, r_2) s.t. $(r_1, r_2) \in E$ and $(r_1, r_2) \notin F$. In words, this is equivalent to the set of all sequences of dice rolls that both have an odd sum and include exactly 0 ones. The set of all such sequences (r_1, r_2) is:

$$EF^c = \{(2,3),(2,5),(3,2),(3,4),(3,6),(4,3),(4,5),(5,2),(5,4),(5,6),(6,3),(6,5)\}$$

v) $EFG = (EF)G = \Big(E \cap F\Big) \cap G =$ the set of all sequences of dice rolls (r_1, r_2) s.t. $(r_1, r_2) \in E$, $(r_1, r_2) \in F$, and $(r_1, r_2) \in G$. In words, this is equivalent to the set of all sequences of dice rolls that have an odd sum, include at least one 1, and that have a sum of exactly 5. The set of all such sequences (r_1, r_2) is:

$$EFG = \{(1,4),(4,1)\}$$

Note: $EFG = FG$ because $EFG = E(FG)$, and all sequences which have a sum of exactly 5 also have an odd sum.

(b) There are 6 possibilities for the first roll and 6 possibilities for the second roll. The rolls are independent, so there are $6^2 = 36$ total possible sequences of two dice rolls. Thus, our sample space is

$$\begin{aligned} S = \{1,2,3,4,5,6\}^2 = \quad & \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6), \\ & (2,1),(2,2),(2,3),(2,4),(2,5),(2,6), \\ & (3,1),(3,2),(3,3),(3,4),(3,5),(3,6), \\ & (4,1),(4,2),(4,3),(4,4),(4,5),(4,6), \\ & (5,1),(5,2),(5,3),(5,4),(5,5),(5,6), \\ & (6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\} \end{aligned}$$

Since we assume the dice are fair, each one of these 36 outcomes is equally likely. Thus, for each of the events, the probability of the event is just

$$\mathbb{P}(event) = \frac{|event|}{|S|} = \frac{|event|}{36}$$

We can calculate the probability of each event using this formula.

1) $\mathbb{P}(E)$: $|E| = 18 \implies$

$$\mathbb{P}(E) = \frac{18}{36} = \frac{1}{2} = 50\%$$

2) $\mathbb{P}(F)$: $|F| = 11 \implies$

$$\mathbb{P}(F) = \frac{11}{36} \approx 30.56\%$$

3) $\mathbb{P}(G)$: $|G| = 4 \implies$

$$\mathbb{P}(G) = \frac{4}{36} = \frac{1}{9} \approx 11.11\%$$

4) $\mathbb{P}(EF)$: $|EF| = 6 \implies$

$$\mathbb{P}(EF) = \frac{6}{36} = \frac{1}{6} \approx 16.67\%$$

5) $\mathbb{P}(E \cup F)$: $|E \cup F| = 23 \implies$

$$\mathbb{P}(E \cup F) = \frac{23}{36} \approx 63.89\%$$

6) $\mathbb{P}(FG)$: $|FG| = 2 \implies$

$$\mathbb{P}(FG) = \frac{2}{36} = \frac{1}{18} \approx 5.56\%$$

7) $\mathbb{P}(EF^c)$: $|EF^c| = 12 \implies$

$$\mathbb{P}(EF^C) = \frac{12}{36} = \frac{1}{3} \approx 33.33\%$$

8) $P(EFG)$: $|EFG| = 2 \implies$

$$P(EFG) = \frac{2}{36} = \frac{1}{18} \approx 5.56\%$$

6. (Ross P2.7) Consider an experiment that consists of determining the type of job—either blue-collar or white-collar—and the political affiliation—Republican, Democratic, or Independent—of the 15 members of an adult soccer team. How many outcomes are

   (a) in the sample space?

   (b) in the event that at least one of the team members is a blue-collar worker?

   (c) in the event that none of the team members considers himself or herself an Independent?

   Solution.
   (a) First, let's define explicitly what we mean by an *outcome* of this experiment.
   Let an *outcome* be any assignment of one of the two job types and one of the three political affiliations to each of the 15 team members.
   For each team member, there are two choices for that member's job type and three choices for that member's political affiliation. These choices are made independently, so, for each team member, there are a total of $2 \cdot 3 = 6$ ways to assign them a job type and a political affiliation. This choice of the assignment for member m_i is independent from the choice of the assignment for member m_j for all $j \neq i$, and there are 15 total team members. Thus, there are a total of 6^{15} ways to assign one of the two job types and one of the three political affiliations to each of the 15 team members. Thus, there are 6^{15} distinct possible outcomes, so the number of outcomes in the sample space is 6^{15}.

   (b) The number of outcomes in which at least one of the team members is a blue-collar worker is equal to the total number of outcomes in the sample space, 6^{15}, minus the total number of outcomes in which *none* of the team members are blue-collar workers.
   If none of the team members are blue-collar workers, then there is only 1 option for the job type of each member (white-collar). This decreases the total number of ways to assign an individual team member one job type and one political affiliation to 3, the number of different political affiliations. The choice of political affiliation is still made independently for each of the 15 team members, so there are 3^{15} possible ways to to assign one job type and one political affiliation to each team member such that *none* of the members are blue-collar workers.
   Thus the total number of outcomes in which *none* of the team members are blue-collar workers is 3^{15}.
   Using the aforementioned formula, we can easily calculate that the total number of outcomes in the event that at least one of the team members is a blue-collar workers is

$$6^{15} - 3^{15}$$

   (c) If none of the team members considers himself or herself an Independent, then, for each team member, there are only 2 options for political affiliation (Republican or Democratic). For each team member, there are still 2 options for job type (blue-collar or white-collar). Thus, for each team member, there are $2 * 2 = 4$ ways to assign that member one job type and one political affiliation if none of the team members is an Independent. The assignment of job type and political affiliation is still independent for each of the 15 team members, so there are

$$4^{15}$$

   total outcomes in the event that none of the team members considers himself or herself an Independent.

7. (Ross P2.9) A retail establishment accepts either the American Express or the VISA credit card. A total of 24 percent of its customers carry an American Express card, 61 percent carry a VISA card, and 11 percent carry both cards. What percentage of its customers carry a credit card that the establishment will accept?

Solution.
Let A = the event that one of the establishment's customers carries an American Express card.
Let V = the event that one of the establishment's customers carries a VISA credit card.
Then the percentage of the establishment's customers that carry a credit card it will accept is $\mathbb{P}(A \cup V)$.

Note: $\mathbb{P}(A \cup V) = \mathbb{P}(A) + \mathbb{P}(V) - \mathbb{P}(A \cap V)$. This is because summing the probabilities of V and A accounts for all outcomes in $A \cup V$, but it double counts all customers with both VISA and American Express cards.

Since 24% of the establishment's customers carry an American Express card, we know

$$\mathbb{P}(A) = 24\%$$

Since 61% of the establishment's customers carry a VISA card, we know

$$\mathbb{P}(V) = 61\%$$

Since 11% of the establishment's customers carry both an American Express and a VISA card, we know

$$\mathbb{P}(A \cap V) = 11\%$$

By applying the aforementioned formula, we can easily calulate that

$$\mathbb{P}(A \cup V) = \mathbb{P}(A) + \mathbb{P}(V) - \mathbb{P}(A \cap V) = 24\% + 61\% - 11\% = 74\%$$

Thus, the percentage of the establishment's customers that carry a credit card that it will accept it 74%.

8. (Ross TE2.10) Prove that

$$P(E \cup F \cup G) = P(E) + P(F) + P(G) - P(E^c FG) - P(EF^c G) - P(EFG^c) - 2P(EFG).$$

Solution. Consider the events $E-(F\cup G)$, $F-(E\cup G)$, $G-(E\cup F)$, E^cFG, EF^cG, EFG^c, and EFG.
$E-(F\cup G)$ refers to all outcomes in E but *not* in F and *not* in G.
$F-(E\cup G)$ refers to all outcomes in F but *not* in E and *not* in G.
$G-(E\cup F)$ refers to all outcomes in G but *not* in E and *not* in F.
E^cFG refers to all outcomes in F and in G but *not* in E.
EF^cG refers to all outcomes in E and in G but *not* in F.
EFG^c refers to all outcomes in E and in F but *not* in G.
Finally, EFG refers to all outcomes in E and in F and in G. Thus, $E-(F\cup G)$, $F-(E\cup G)$, $G-(E\cup F)$, E^cFG, EF^cG, EFG^c, and EFG are all mutually disjoint, and

$$\Big(E-(F\cup G)\Big)\cup\Big(F-(E\cup G)\Big)\cup\Big(G-(E\cup F)\Big)\cup\Big(E^cFG\Big)\cup\Big(EF^cG\Big)\cup\Big(EFG^c\Big)\cup EFG = E\cup F\cup G$$

Since the events are mutually disjoint, we have

$$\begin{aligned}\mathbb{P}(E\cup F\cup G) = \quad & \mathbb{P}(E-(F\cup G)) + \mathbb{P}(F-(E\cup G)) + \mathbb{P}(G-(E\cup F)\\ & + \mathbb{P}(E^cFG) + \mathbb{P}(EF^cG) + \mathbb{P}(EFG^c) + \mathbb{P}(EFG)\end{aligned}$$

Note: Since $(F\cup G)E$ refers to all outcomes in E and either in F or in G, or both, we know that

$$E = \Big(E-(F\cup G)\Big)\cup(F\cup G)E$$

Also, $(F\cup G)E$ and $E-(F\cup G)$ are mutually disjoint, so

$$\mathbb{P}(E) = \mathbb{P}(E-(F\cup G)) + \mathbb{P}((F\cup G)E)$$

Thus,

$$\mathbb{P}(E-(F\cup G)) = \mathbb{P}(E) - \mathbb{P}((F\cup G)E)$$

By symmetry, we also know $\begin{cases}\mathbb{P}(F-(E\cup G)) = \mathbb{P}(F) - \mathbb{P}(E\cup G)F\\ \mathbb{P}(G-(E\cup F)) = \mathbb{P}(G) - \mathbb{P}(E\cup F)G\end{cases}$

Thus,

$$\begin{aligned}\mathbb{P}(E\cup F\cup G) = \quad & \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}((F\cup G)E) - \mathbb{P}((E\cup G)F) - \mathbb{P}((E\cup F)G)\\ & + \mathbb{P}(E^cFG) + \mathbb{P}(EF^cG) + \mathbb{P}(EFG^c) + \mathbb{P}(EFG)\end{aligned}$$

Note: Since EFG refers to all outcomes in E, in F, and in G, EF^cG refers to all outcomes in E, *not* in F, and in G, and EFG^c refers to all outcomes in E, in F, and *not* in G, we know

$$(F\cup G)E = EFG\cup EF^cG\cup EFG^c$$

Since EFG, EF^cG, and EFG^c are mutually disjoint, we know

$$\mathbb{P}((F\cup G)E) = \mathbb{P}(EFG) + \mathbb{P}(EF^cG) + \mathbb{P}(EFG^c)$$

Multiplying both sides by -1, we find

$$-\mathbb{P}((F\cup G)E) = -\mathbb{P}(EFG) - \mathbb{P}(EF^cG) - \mathbb{P}(EFG^c)$$

By symmetry, we also know $\begin{cases} -\mathbb{P}((E \cup G)F) = -\mathbb{P}(EFG) - \mathbb{P}(E^cFG) - \mathbb{P}(EFG^c) \\ -\mathbb{P}((E \cup F)G) = -\mathbb{P}(EFG) - \mathbb{P}(E^cFG) - \mathbb{P}(EF^cG) \end{cases}$

Plugging this into the equation for $\mathbb{P}(E \cup F \cup G)$, we find

$$\begin{aligned} \mathbb{P}(E \cup F \cup G) = \quad & \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}((F \cup G)E) - \mathbb{P}((E \cup G)F) - \mathbb{P}((E \cup F)G) \\ & + \mathbb{P}(E^cFG) + \mathbb{P}(EF^cG) + \mathbb{P}(EFG^c) + \mathbb{P}(EFG) \\ = \quad & \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(EFG) - \mathbb{P}(EF^cG) - \mathbb{P}(EFG^c) - \mathbb{P}(EFG) \\ & - \mathbb{P}(E^cFG) - \mathbb{P}(EFG^c) - \mathbb{P}(EFG) - \mathbb{P}(E^cFG) - \mathbb{P}(EF^cG) \\ & + \mathbb{P}(E^cFG) + \mathbb{P}(EF^cG) + \mathbb{P}(EFG^c) + \mathbb{P}(EFG) \\ = \quad & \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(E^cFG) - \mathbb{P}(EF^cG) - \mathbb{P}(EFG^c) - 2\mathbb{P}(EFG) \end{aligned}$$

as required. This completes the proof that

$$\mathbb{P}(E \cup F \cup G) = \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(E^cFG) - \mathbb{P}(EF^cG) - \mathbb{P}(EFG^c) - 2\mathbb{P}(EFG).$$

9. (Ross TE2.12) Show that the probability that exactly one of the events E or F occurs equals $P(E) + P(F) - 2P(EF)$.
Let $X =$ exactly one of the events E or F occurs.
We want to prove that $\mathbb{P}(X) = \mathbb{P}(E) + \mathbb{P}(F) - 2\mathbb{P}(EF)$.
Note: $X \cup EF = (E \cup F)$. Since X and EF are mutually disjoint, it directly follows that

$$\mathbb{P}(X) + \mathbb{P}(EF) = \mathbb{P}(E \cup F)$$

Subtracting $\mathbb{P}(EF)$ from both sides, we find

$$\mathbb{P}(X) = \mathbb{P}(E \cup F) - \mathbb{P}(EF)$$

We already know that

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(EF)$$

Plugging this into our equation for $\mathbb{P}(X)$, we find

$$\mathbb{P}(X) = \mathbb{P}(E \cup F) - \mathbb{P}(EF) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(EF) - \mathbb{P}(EF) = \mathbb{P}(E) + \mathbb{P}(F) - 2\mathbb{P}(EF)$$

as required. This completes the proof that the probability that exactly one of the events E or F occurs equals $\mathbb{P}(E) + \mathbb{P}(F) - 2\mathbb{P}(EF)$.

10. (Ross P2.29) An urn contains n white and m black balls, where n and m are positive numbers.

(a) If two balls are randomly withdrawn, what is the probability that they are the same color?

(b) If a ball is randomly withdrawn and then replaced before the second one is drawn, what is the probability that the withdrawn balls are the same color?

(c) Show that the probability in part (b) is always larger than the one in part (a).

Solution.
(a) There are $n+m$ total balls in the urn, and we want to select 2. Since the balls are withdrawn simultaneously, there is no order in which they are withdrawn. Thus, there are $\binom{n+m}{2}$ equally likely ways to select two balls simultaneously from the urn. Thus, the size of the sample space is $|S| = \binom{n+m}{2}$

If the two balls are the same color, then they are either both white or both black.

1) Both balls are white:
We need to select 2 white balls from the n white balls in the urn. Since the balls are withdrawn simultaneously, there is no order in which they are withdrawn. Thus, there are a total of $\binom{n}{2}$ equally likely ways to simultaneously select two white balls from the urn.

2) Both balls are black:
We need to select 2 black balls from the m white balls in the urn. Since the balls are withdrawn simultaneously, there is no order in which they are withdrawn. Thus, there are a total of $\binom{m}{2}$ equally likely ways to simultaneously select two black balls from the urn.

Thus, the total number of ways to simultaneously draw 2 balls of the same color from the urn is $\binom{n}{2} + \binom{m}{2}$.
Let $E =$ the event that the two balls are the same color (without replacement).
Then, since all outcomes are equally likely,

$$\mathbb{P}(E) = \frac{|E|}{|S|} = \frac{\binom{n}{2} + \binom{m}{2}}{\binom{n+m}{2}} = \frac{\frac{n(n-1)}{2} + \frac{m(m-1)}{2}}{\frac{(m+n)(m+n-1)}{2}} = \frac{n(n-1) + m(m-1)}{(m+n)(m+n-1)}$$

Therefore, the probability that both of the balls are the same color (without replacement) is

$$\mathbb{P}(E) = \frac{n(n-1) + m(m-1)}{(m+n)(m+n-1)}$$

(b) There are still $n+m$ total balls in the urn, but now the balls are drawn one at a time, with replacement. Thus, we have $n+m$ options for the first ball we choose, and we still have $n+m$ options for the second ball we choose. This results in a total of $(n+m)^2$ equally likely ways to draw one ball, replace it, then draw another. Thus, the size of the sample space is $|S'| = (n+m)^2$

Once again, if the two balls are the same color, then they are either both white or both black.

1. Both balls are white:
We have n ways to select a first white ball. Since we replace our first ball before choosing our second, we also have n ways to select our second white ball. Thus, there are a total of n^2 equally likely ways to draw one white ball, replace it, then draw another white ball.

2. Both balls are black:
We have m ways to select the first black ball. Since we replace our first ball before choosing our second, we also have m ways to select our second black ball. Thus, there are exactly m^2 equally likely ways to select one black ball, replace it, then select another black ball.

Thus, the total number of ways to draw one ball, replace it, then draw another of the same color is $n^2 + m^2$.
Let $F =$ the event that the two balls are the same color (with replacement).
Then, since all outcomes are equally likely,

$$\mathbb{P}(F) = \frac{|F|}{|S'|} = \frac{n^2 + m^2}{(n+m)^2}$$

Therefore, the probability that both of the balls are the same color (with replacement) is

$$\mathbb{P}(F) = \frac{n^2 + m^2}{(n+m)^2}$$

(c) *Note:* It suffices to show that $\mathbb{P}(F) > \mathbb{P}(E)$.
Thus, it also suffices to show $\mathbb{P}(F) - \mathbb{P}(E) > 0$.

$$\mathbb{P}(F) - \mathbb{P}(E) = \frac{n^2 + m^2}{(n+m)^2} - \frac{n(n-1) + m(m-1)}{(m+n)(m+n-1)}$$

We can now rewrite both fractions in terms of the common denominator $(n+m)^2(n+m-1)$ to combine the fractions into a single term as follows:

$$\begin{aligned} \frac{n^2 + m^2}{(n+m)^2} - \frac{n(n-1) + m(m-1)}{(m+n)(m+n-1)} &= \frac{(n^2 + m^2)(n+m-1)}{(n+m)^2(n+m-1)} - \frac{(n(n-1) + m(m-1))(m+n)}{(n+m)^2(n+m-1)} \\ &= \frac{(n^2 + m^2)(n+m-1) - (n(n-1) + m(m-1))(m+n)}{(n+m)^2(n+m-1)} \end{aligned}$$

Expanding and simplifying the numerator, we obtain

$$\begin{aligned} & \frac{n^2 + m^2}{(n+m)^2} - \frac{n(n-1) + m(m-1)}{(m+n)(m+n-1)} \\ = \quad & \frac{(n^2 + m^2)(n+m-1) - (n(n-1) + m(m-1))(m+n)}{(n+m)^2(n+m-1)} \\ = \quad & \frac{(n^3 + n^2m - n^2 + m^3 + m^2n - m^2) - (n^3 - n^2 + n^2m - nm + nm^2 + m^3 - nm - m^2)}{(n+m)^2(n+m-1)} \\ = \quad & \frac{n^3 + n^2m - n^2 + m^3 + m^2n - m^2 - n^3 + n^2 - n^2m + nm - nm^2 - m^3 + nm + m^2}{(n+m)^2(n+m-1)} \\ = \quad & \frac{2nm}{(n+m)^2(n+m-1)} > 0 \end{aligned}$$

as required. This completes the proof that

$$P(F) > P(E)$$

.

11. (a) Give a probabilistic proof of the geometric series identity in the form

$$(1-p)\sum_{i=0}^{\infty} p^i = 1$$

for any $0 \leq p \leq 1$.

(b) Give a probabilistic proof of the related identity

$$\sum_{i=0}^{\infty}(i+1)p^i = \frac{1}{(1-p)^2}.$$

Solution.
(a) Consider an experiment in which we roll a die until the value v appears once, at which point the experiment stops. For each die roll, let $p =$ the probability that we *do not* roll a v. Then $0 \leq p \leq 1$ as required, and $1 - p =$ the probability that we *do* roll a v on any given roll of the die.

We will now argue that both sides of the identity

$$(1-p)\sum_{i=0}^{\infty} p^i = 1 \qquad (1)$$

are equal to the probability of the sample space, $\mathbb{P}(S)$, in our experiment.

Right-Hand Side: $\mathbb{P}(S)$ is trivially equal to the Right-Hand Side of (1), as the second axiom of probability guarantees that

$$\mathbb{P}(S) = 1$$

for any experiment.

Left-Hand Side: Let's try to find the a way to express our sample space S. We can do this by counting the number of rolls in possible outcomes of the experiment. If v first appears on on the first roll, then our experiment takes exactly 1 roll. If v first appears on the second roll, then our experiment takes exactly 2 rolls. Similarly, for any $i \in \mathbb{N}$, if v first appears on the i'th roll, then our experiment takes exactly i rolls. Thus, there are outcomes in our sample space that take exactly i rolls for all $i \in \mathbb{N}$. Additionally, there is the possibility that v never appears, in which case we must keep rolling the die infinitely. Thus, there is also an outcome in our sample space that takes infinitely many rolls. Let $S_i =$ the event that our experiment takes exactly i rolls, and let $X =$ the event that v never appears in our experiment. Then we can think of the sample space as the union between X and S_i, for all $i : 1 \to \infty$. That is,

$$S = \Big(\bigcup_{i=1}^{\infty} S_i\Big) \cup X$$

Note: For all outcomes s, if s takes exactly i rolls, it is impossible for s to take exactly j rolls, for all $j \neq i$. Thus,

$$s \in S_i \implies s \notin S_j$$

for all $j \neq i$.
Also,

$$s \in X \implies s \notin S_i$$

for all $i \in N$. Therefore, $\{X, S_1, S_2, \cdots\}$ is a mutually disjoint set, so we know

$$\mathbb{P}(S) = \mathbb{P}(X) + \sum_{i=1}^{\infty} \mathbb{P}(S_i)$$

Let's try to find a probabilistic formula for $\mathbb{P}(X)$. If $s \in X$, then we know we have infinitely many dice rolls, all of which are *not* the value v. The probability of *not* rolling v on any given roll, independent of the roll number, is p, so the probability of an infinite sequence of rolls that all *aren't* v is

$$\mathbb{P}(X) = \lim_{i \to \infty} p^i = 0$$

for all $p < 1$. This makes sense, as the probability of rolling infinitely many times without getting a v should be zero unless the probability of not getting a v is 100 %.
Thus,

$$\mathbb{P}(S) = 0 + \sum_{i=1}^{\infty} \mathbb{P}(S_i) = \sum_{i=1}^{\infty} \mathbb{P}(S_i)$$

Now, let's try to find a probabilistic formula for $\mathbb{P}(S_i)$. If an outcome $s \in S_i$, then v first appeared on the i'th roll of that outcome of the experiment. Thus, v did not appear during the first $i-1$ rolls of the experiment. On any given roll, the probability of not rolling v is p. Each of these probabilities is independent, so the probability of *not* rolling a v on all of the first $i-1$ rolls is p^{i-1}. On any given roll, the probability of rolling a v is $1-p$, and this probability is also independent of the roll number. Thus, the probability of *not* rolling a v on all of the first $i-1$ rolls and then rolling a v on the i'th roll is

$$\mathbb{P}(S_i) = (1-p)p^{i-1}$$

This is true for all i from $1 \to \infty$, so

$$\mathbb{P}(S) = \sum_{i=1}^{\infty} \mathbb{P}(S_i) = \sum_{i=1}^{\infty} (1-p)p^{i-1}$$

i takes values from $1 \to \infty$, so $i-1$ takes values from $0 \to \infty$, so we can rewrite the sum as

$$\sum_{i=1}^{\infty} (1-p)p^{i-1} = \sum_{i=0}^{\infty} (1-p)p^i$$

Now, since $(1-p)$ does not depend on i, we can pull it out of the sum, applying the distributive property of addition, to find

$$\sum_{i=0}^{\infty} (1-p)p^i = (1-p)\sum_{i=0}^{\infty} p^i$$

which is exactly the Left-Hand Side of the identity we want to prove. Thus, we have shown that

$$(1-p)\sum_{i=0}^{\infty} p^i = \mathbb{P}(S) = 1$$

which completes the probabilistic proof that

$$(1-p)\sum_{i=0}^{\infty} p^i = 1$$

(b) *Note:* It suffices to show

$$(1-p)^2 \sum_{i=0}^{\infty} (i+1)p^i = 1 \qquad (2)$$

We will apply a very similar probabilistic argument, with a slightly different setup. This time, consider an experiment in which we roll a die until the value v appears twice, at which point the experiment stops. For each die roll, let $p =$ the probability that we *do not* roll a v. Then $0 \leq p \leq 1$ as required, and

$1 - p =$ the probability that we *do* roll a v on any given roll of the die. We will now argue that both sides of the identity

$$(1-p)^2 \sum_{i=0}^{\infty} (i+1)p^i = 1 \qquad (2)$$

are equal to the probability of the sample space, $\mathbb{P}(S)$, in our experiment.

Right-Hand Side: $\mathbb{P}(S)$ is trivially equal to the Right-Hand Side of (2), as the second axiom of probability guarantees that

$$\mathbb{P}(S) = 1$$

for any experiment.

Left-Hand Side: Once again, let's try to find the a way to express our sample space S. v cannot appear a second time by the first roll, so we know our experiment takes ≥ 2 rolls to finish. v could appear for a second time on the second roll, and v could appear for a second time on the i'th roll, for all $i \in \mathbb{N}$ s.t. $i \geq 2$. However, v could also never appear for a second time, meaning we must continue to roll the die infinitely. Thus, the number of rolls taken by different outcomes of the experiment ranges from $i : 2 \to \infty$. Let $S_i =$ the event that the experiment takes exactly i rolls, and let $X =$ the event that v never appears twice during the experiment. Just like in part (a)

$$S = \Big(\bigcup_{i=2}^{\infty} S_i\Big) \cup X$$

Once again, all S_i are mutually disjoint, so

$$\mathbb{P}(S) = \mathbb{P}(X) + \sum_{i=2}^{\infty} \mathbb{P}(S_i)$$

Let's try to find a probabilistic formula for $\mathbb{P}(X)$. $x \in X$ implies there are infinitely rolls in outcome x, at which point v never appears twice. Thus, we could either have v appear once, or v never appear in the outcome x. If v appears once, it could have appeared during any of the infinitely many rolls, except the last one. Thus, we have

$$\mathbb{P}(X) = \lim_{i \to \infty} (i-1)(1-p)p^{i-1} = 0$$

for all $p < 1$.
Similarly, if v never appears, then we know all of the infinitely many rolls were *not* v, so all of the infinitely many rolls had probability p, so we have

$$\mathbb{P}(X) = \lim_{i \to \infty} p^i = 0$$

for all $p < 1$.
Thus, just like in part (a), the probability that our experiment takes infinitely many rolls is 0, as long as the probability of rolling a v is nonzero, so we have

$$\mathbb{P}(S) = 0 + \sum_{i=2}^{\infty} \mathbb{P}(S_i) = \sum_{i=2}^{\infty} \mathbb{P}(S_i)$$

Now, let's try to find a probabilistic formula for $\mathbb{P}(S_i)$. On any given roll, there is a p probability of *not* rolling a v, and a $(1-p)$ probability of rolling a v. For $s \in S_i$, exactly 2 of those i rolls were v's, and the remaining $i-2$ rolls were not v's. These probabilities are independent for each roll, so, for any individual outcome $s \in S_i$,

$$\mathbb{P}(s) = (1-p)^2 p^{i-2}$$

Now, lets count how many distinct $s \in S_i$. Since the experiment only stops when v appears for a second time, we know the last value rolled was v (assuming we didn't take infinitely many rolls). Also, we know v appeared in exactly 1 of the first $i-1$ rolls. This gives us $i-1$ different choices for the location of the first v, each of which results in a distinct $s \in S_i$. Thus, there are $i-1$ distinct outcomes in S_i, each of which has a probability of $(1-p)^2 p^{i-2}$. Thus, the total probability of each S_i is

$$\mathbb{P}(S_i) = (1-p)^2(i-1)p^{i-2}$$

Plugging this into the equation for $\mathbb{P}(S)$, we find

$$\mathbb{P}(S) = \sum_{i=2}^{\infty} \mathbb{P}(S_i) = \sum_{i=2}^{\infty} (1-p)^2(i-1)p^{i-2}$$

Now, since $(1-p)$ does not depend on i, we can pull it out of the sum, applying the distributive property of addition, to find

$$\sum_{i=2}^{\infty} (1-p)^2(i-1)p^{i-2} = (1-p)^2 \sum_{i=2}^{\infty} (i-1)p^{i-2}$$

i takes values from $2 \to \infty$, so $i-2$ takes values from $0 \to \infty$, so we can rewrite the sum as

$$(1-p)^2 \sum_{i=2}^{\infty} (i-1)p^{i-2} = (1-p)^2 \sum_{i=0}^{\infty} (i+1)p^{i}$$

Thus, we have shown that

$$(1-p)^2 \sum_{i=0}^{\infty} (i+1)p^{i} = \mathbb{P}(S) = 1$$

which completes the probabilistic proof that

$$\sum_{i=0}^{\infty} (i+1)p^{i} = \frac{1}{(1-p)^2}$$

Assignment 5

Math 407 (Swanson) – Spring 2023
Homework 1
Due Friday 1/13, 11:59pm

Name: Emerson Kahle Section: 39981

- You must upload your solutions to Gradescope as **one single, high-quality PDF**. You can convert paper-based work to a high-quality PDF using a scanning app for mobile devices, such as Adobe Scan (free, available for iOS and Android, can do multiple pages) or many others. If necessary, you can combine or merge multiple PDF's into a single PDF using a variety of services, such as Adobe Acrobat's cloud-based merge tool.

- After you upload, you must match each question with its corresponding page using Gradescope's interface. This allows graders to spend more time giving you feedback instead of hunting through submissions.
- Answers without supporting work will receive no credit. Show your work.
- You are encouraged to work together on homework, but **you must write up your solutions separately in your own words**. Copying from your fellow students or other sources is a serious academic integrity violation. In particular, you may not use "tutoring" services which simply provide answers.
- You are encouraged to typeset your solutions in LaTeX. Source code has been provided on Blackboard. Overleaf is a popular cloud-based editor.
- Problem numbers refer to the course textbook, though the problems may have been modified significantly.

1. (Ross P2.21) A small community organization consists of 20 families, of which 4 have one child, 8 have two children, 5 have three children, 2 have four children, and 1 has five children.

   (a) If one of these families is chosen at random, what is the probability it has i children, $i = 1, 2, 3, 4, 5$?

   (b) If one of the children is randomly chosen, what is the probability that child comes from a family having i children, $i = 1, 2, 3, 4, 5$?

   Solution.
   (a) Since we are choosing from the families at random, there is an equal probability of selecting each of the 20 families. There are 20 total families, so the size of our sample space S is $|S| = 20$ However, the probability that the family we select has i children depends on the value of i. Let E_i = the event that the chosen family has i children.

   (i) $i = 1$: There are 4 families that that have exactly one child, and we have an equal probability of selecting each family, so the probability that the selected family has exactly 1 child is

   $$\mathbb{P}(E_1) = \frac{|E_1|}{|S|} = \frac{4}{20} = \frac{1}{5} = 20\%$$

   (ii) $i = 2$: There are 8 families that have exactly two children, and we have an equal probability of selecting each family, so the probability that the selected family has exactly two children is

   $$\mathbb{P}(E_2) = \frac{|E_2|}{|S|} = \frac{8}{20} = \frac{2}{5} = 40\%$$

   (iii) $i = 3$: There are 5 families that have exactly three children, and we have an equal probability of selecting each family, so the probability that the selected family has exactly three children is

   $$\mathbb{P}(E_3) = \frac{|E_3|}{|S|} = \frac{5}{20} = \frac{1}{4} = 25\%$$

   (iv) $i = 4$: There are 2 families that have exactly four children, and we have an equal probability of selecting each family, so the probability that the selected family has exactly four children is

   $$\mathbb{P}(E_4) = \frac{|E_4|}{|S|} = \frac{2}{20} = \frac{1}{10} = 10\%$$

   (v) $i = 5$: There is only 1 family that has exactly five children, and we have an equal probability of selecting each family, so the probability that the selected family has exactly five children is

   $$\mathbb{P}(E_5) = \frac{|E_5|}{|S|} = \frac{1}{20} = 5\%$$

   Note: As expected, since the E_i's are all mutually disjoint, and the selected family must have exactly 1, 2, 3, 4, or 5 children, we find that

   $$\mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3) + \mathbb{P}(E_4) + \mathbb{P}(E_5) = \frac{4}{20} + \frac{8}{20} + \frac{5}{20} + \frac{2}{20} + \frac{1}{20} = \frac{20}{20} = 1$$

   (b) Similarly, since we are selecting from the children at random, there is an equal probability of selecting each child. Therefore, we need to count up all the children to find a size for our sample space. Then, we can use the same process from part (a) to calculate the probabilities for each i, this time with the event F_i = the selected child comes from a family with exactly i children. Adding up the children from all 20 families, we find there are exactly

   $$4 \cdot 1 + 8 \cdot 2 + 5 \cdot 3 + 2 \cdot 4 + 1 \cdot 5 = 4 + 16 + 15 + 8 + 5 = 48$$

   children between the 20 families. Therefore, the size of our sample space, S', is $|S'| = 48$. The probability that the selected child belongs to a family with exactly i children depends on the value of i.

(i) $i = 1$: There are 4 families with exactly one child, for a total of $4 * 1$ children that belong to such a family. Since we have an equal probability of selecting each child, we know the probability that the selected child belongs to a family with exactly one child is

$$\mathbb{P}(F_1) = \frac{|F_1|}{|S'|} = \frac{4}{48} = \frac{1}{12} \approx 8.33\%$$

(ii) $i = 2$: There are 8 families with exactly two children, for a total of $8 * 2 = 16$ children that belong to such a family. Since we have an equal probability of selecting each child, we know the probability that the selected child belongs to a family with exactly two children is

$$\mathbb{P}(F_2) = \frac{|F_2|}{|S'|} = \frac{16}{48} = \frac{1}{3} \approx 33.33\%$$

(iii) $i = 3$: There are 5 families with exactly three children, for a total of $5 * 3 = 15$ children that belong to such a family. Since we have an equal probability of selecting each child, we know the probability that the selected child belongs to a family with exactly three children is

$$\mathbb{P}(F_3) = \frac{|F_3|}{|S'|} = \frac{15}{48} = \frac{5}{16} = 31.25\%$$

(iv) $i = 4$: There are 2 families with exactly four children, for a total of $4 * 2 = 8$ children that belong to such a family. Since we have an equal probability of selecting each child, we know the probability that the selected child belongs to a family with exactly four children is

$$\mathbb{P}(F_4) = \frac{|F_4|}{|S'|} = \frac{8}{48} = \frac{1}{6} \approx 16.67\%$$

(v) $i = 5$: There is only 1 family with exactly five children, for a total of $1 \cdot 5 = 5$ children that belong to such a family. Since we have an equal probability of selecting each child, we know the probability that the selected child belongs to a family with exactly five children is

$$\mathbb{P}(F_5) = \frac{|F_5|}{|S'|} = \frac{5}{48} \approx 10.42\%$$

Note: As expected, since the F_i's are all mutually disjoint, and the selected child must belong to a family with 1, 2, 3, 4, or 5 children, we find that

$$\mathbb{P}(F_1) + \mathbb{P}(F_2) + \mathbb{P}(F_3) + \mathbb{P}(F_4) + \mathbb{P}(F_5) = \frac{4}{48} + \frac{16}{48} + \frac{15}{48} + \frac{8}{48} + \frac{5}{48} = \frac{48}{48} = 1$$

2. (Ross P2.25) A pair of dice is rolled until a sum of either 5 or 7 appears. Find the probability that a 5 occurs first. *Hint:* Let E_n denote the event that a 5 occurs on the nth roll and no 5 or 7 occurs on the first $n-1$ rolls. Compute $P(E_n)$ and argue that $\sum_{n=1}^{\infty} P(E_n)$ is the desired probability.

Solution.
Let S_5 = the event that the sum of a given pair of dice rolls equals 5.
Let S_7 = the event that the sum of a given pair of dice rolls equals 7.
The sample space for a pair of dice rolls is

$$\begin{aligned} S = \{1,2,3,4,5,6\}^2 = \quad & \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ & (2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ & (3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ & (4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ & (5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\ & (6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\} \end{aligned}$$

Each of these outcomes is equally likely, and only 4 outcomes, $(1,4),(2,3),(3,2),(4,1)$ have sums of 5. Therefore, $S_5 = \{(1,4),(2,3),(3,2),(4,1)\}$, and the probability that the sum of a given pair of dice rolls equals 5 is

$$\mathbb{P}(S_5) = \frac{|S_5|}{|S|} = \frac{4}{36} = \frac{1}{9} \approx 11.11\%$$

Similarly, only 6 outcomes, $(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)$ have sums of 7. Therefore, $S_7 = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$ and the probability that the sum of a given pair of dice rolls equals 7 is

$$\mathbb{P}(S_7) = \frac{|S_7|}{|S|} = \frac{6}{36} = \frac{1}{6} \approx 16.67\%$$

If a given pair of dice rolls has a sum of 7, it cannot have a sum of 5, and vice-versa. Therefore, S_7 and S_5 are mutually disjoint, so the probability that a given pair of dice rolls has a sum of 5 *or* 7 is

$$\mathbb{P}(S_5 \cup S_7) = \mathbb{P}(S_5) + \mathbb{P}(S_7) = \frac{4+6}{36} = \frac{10}{36} = \frac{5}{18} \approx 27.78\%$$

Let S^* = the event that a given pair of dice rolls has a sum that is neither 5 nor 7. Then

$$S^* = (S_5 \cup S_7)^c \implies \mathbb{P}(S^*) = 1 - \mathbb{P}(S_5 \cup S_7) = 1 - \frac{5}{18} = \frac{13}{18} \approx 72.22\%$$

Now, let E_n = the event that a 5 occurs on the nth roll and no 5 or 7 occurs on the first $n-1$ rolls. We want to find $\mathbb{P}(E_n)$ in terms of n. For any outcome $e \in E_n$, each of the n pairs of dice rolls occur independently, so we can multiply their corresponding probabilities together to get $\mathbb{P}(E_n)$. Each of the first $n-1$ pairs of dice rolls has a sum that is neither 5 nor 7. Therefore, each of the first $n-1$ pairs of dice rolls have a probability of

$$\mathbb{P}(S^*) = \frac{13}{18}$$

The nth dice roll must have a sum of 5, so its probability is

$$\mathbb{P}(S_5) = \frac{1}{9}$$

Therefore, the total probability that the first $n-1$ pairs of rolls never sum to 5 or 7 and the nth pair of rolls has a sum of 5 is

$$\mathbb{P}(E_n) = (\frac{13}{18})^{n-1}(\frac{1}{9})$$

Note: For all $e \in E_n$, $e \notin E_j$ for all $j \neq n$.
Proof. Assume to the contrary that $\exists\ e \in E_n$ s.t. $e \in E_j$ but $j \neq n$. Without loss of generality, assume

$n > j$. Then $e \in E_n$ implies that none of the first $n-1$ pairs of dice rolls had a sum of 5. However, one of the first $n-1$ pairs of dice rolls was the jth pair, and $e \in E_j$ implies that a sum of 5 appeared on the jth pair of dice rolls. This contradiction completes the proof.
From this proof, we know that all E_n are mutually disjoint.

We want to find the probability that a sum of 5 appears first. Consider any such outcome. Since the experiment ends after the first 5 or 7 appears, we know that only one pair of dice rolls (the last one) had a sum of 5. The 5 could first appear on the first pair of rolls, or the second, or the third, all the way to the nth, for all $n \in \mathbb{N}$. Since the 5 appears only once in any of these outcomes, we know it cannot appear on two different pairs of dice rolls. Thus, if we calculate the probability of the 5 appearing on the nth roll for all $n \in \mathbb{N}$, we can add these probabilities together to find the total probability that a sum of 5 appears before a sum of 7 *(Observation 1)*.
To do so, we just need to calculate the probability of the 5 appearing first on the nth roll in terms of n.

Note: The probability of the 5 appearing first on the nth roll is equal to $\mathbb{P}(E_n)$.
Proof. It suffices to show that for all $e \in E_n$, a 5 appears first on the nth roll of e, and that for all outcomes o in which a 5 appears first on the nth roll of o, $o \in E_n$.
First, we will show the former. By the definition of E_n, for all $e \in E_n$, we know a 5 occurs on the nth pair of rolls, and we know that no 5 nor 7 appears during the first $n-1$ pairs of rolls. Thus, for all $e \in E_n$, e is an outcome in which 5 appears first on the nth roll.
Now, we can show the latter. For all such o, a 5 appears first on the nth pair of rolls. Since the experiment stops as soon as a 5 or 7 appears, we know that no 5 nor 7 appeared during the first $n-1$ pairs of rolls in o. Therefore, for all o in which 5 appears first on the nth roll, o satisfies both of the conditions for E_n, so $o \in E_n$.
Thus, we have shown that $\{5 \text{ first appears on the } n\text{th roll}\} = E_n$ for all $n \in \mathbb{N}$, which completes the proof that

$$\mathbb{P}(5 \text{ first appears on the } n\text{th roll}) = \mathbb{P}(E_n) = (\frac{13}{18})^{n-1}(\frac{1}{9})$$

for all $n \in \mathbb{N}$.

Combining this result with *Observation 1*, we find that the total probability that a sum of 5 appears before a sum of 7 is

$$\sum_{n=1}^{\infty} \mathbb{P}(E_n) = \sum_{n=1}^{\infty} (\frac{13}{18})^{n-1}(\frac{1}{9}) = \frac{1}{9}\sum_{n=1}^{\infty} (\frac{13}{18})^{n-1} = \frac{1}{9}\sum_{n=0}^{\infty} (\frac{13}{18})^{n} \qquad (1)$$

Now, $\sum_{n=0}^{\infty}(\frac{13}{18})^n$ is a geometric series, so we can apply the geometric series identity for $|r| < 1$ to find

$$\sum_{n=0}^{\infty} (\frac{13}{18})^{n} = \frac{1}{1-\frac{13}{18}} = \frac{1}{\frac{5}{18}} = \frac{18}{5}$$

Combining this with (1), we find that the total probability that a sum of 5 appears before a sum of 7 in our experiment is

$$\frac{1}{9}\sum_{n=0}^{\infty} (\frac{13}{18})^{n} = \frac{1}{9}\frac{18}{5} = \frac{2}{5} = 40\%$$

3. (Ross P2.34) The second Earl of Yarborough is reported to have bet at odds of 1000 to 1 that a bridge hand of 13 cards would contain at least one card that is ten or higher. (By *ten or higher* we mean that a card is either a ten, a jack, a queen, a king, or an ace.) Nowadays, we call a ahnd that has no cards higher than 9 *a Yarborough*. What is the probability that a randomly selected bridge hand is a Yarborough?

Solution.
Note that each bridge hand is equally likely, and there are $\binom{52}{13}$ possible bridge hands. Thus, the size of our sample space, S, is $|S| = \binom{52}{13}$.
Now, we just need to count the number of bridge hands that are *Yarboroughs*. Let Y = a randomly selected bridge hand is a Yarborough. There are 4 tens, 4 jacks, 4 queens, 4 kings, and 4 aces in standard deck. Therefore, there are $4+4+4+4+4 = 4 \cdot 5 = 20$ cards which would prevent a hand from being a *Yarborough*. Since there are 52 cards in a standard deck, there are $52-20 = 32$ cards which we can select from to construct a hand which *is* a *Yarborough*. We need to select 13 cards from these 32 cards, and the permutation of the cards does not matter, so there are $\binom{32}{13}$ equally likely ways to do this. Therefore, out the $\binom{52}{13}$ equally likely bridge hands, exactly $\binom{32}{13}$ of them are *Yarboroughs*. Thus, $|Y| = \binom{32}{13}$, so the probability that a randomly selected bridge hand is a *Yarborough* is

$$\frac{\binom{32}{13}}{\binom{52}{13}} = \frac{\frac{32!}{13! \cdot 19!}}{\frac{52!}{13! \cdot 39!}} = \frac{32!}{13! \cdot 19!}\frac{13! \cdot 39!}{52!} = \frac{39 \cdot 38 \cdots 21 \cdot 20}{52 \cdot 51 \cdots 34 \cdot 33} = \frac{32 \cdot 31 \cdots 21 \cdot 20}{52 \cdot 51 \cdots 41 \cdot 40} \approx 0.05\%$$

4. (Ross P2.37) An instructor gives her class a set of 10 problems with the information that the final exam will consist of a random selection of 5 of them. If a student has figured out how to do 7 of the problems, what is the probability that he or she will answer correctly...

   (a) all 5 problems?

   (b) at least 4 of the problems?

   Solution.
   (a) If the student knows how to do all 5 problems, then they must know how to do the first problem.There are 10 equally likely possibilities for the first problem, of which the student knows how to solve 7. Thus, the probability that the student gets the first question correct is $\frac{7}{10} = 70\%$.
   After getting the first problem correct, there are 9 equally likely possibilities for the second question, of which the student now only knows how to solve 6. Thus, the probability that the student gets the second question correct given that they got the first question correct is $\frac{6}{9}$.
   After getting the first two questions right, there are 8 equally likely possibilities for the third question, of which the student now only knows how to solve 5. Thus, the probability that the student gets the third question right given that they got the first two problems right is $\frac{5}{8}$.
   After getting the first three problems right, there are 7 equally likely possibilities for the fourth problem, of which the student now only knows how to solve 4. Thus, the probability that the student gets the fourth problem right given that they got the first three problems right is $\frac{4}{7}$.
   After getting the first four problems right, there are 6 equally likely possibilities for the fifth problem, of which the student now only knows how to solve 3. Thus, the probability that the student gets the fifth problem right given that they got the first four problems right is $\frac{3}{6}$.
   Multiplying these probabilities together, we find that the total probability that the student answers all 5 problems correctly is

$$\frac{7}{10}\frac{6}{9}\frac{5}{8}\frac{4}{7}\frac{3}{6} = \frac{1}{2}\frac{2}{1}\frac{1}{2}\frac{1}{1}\frac{1}{6} = \frac{2}{24} = \frac{1}{12} \approx 8.33\%$$

   Alternatively, we could note that the number of ways to choose the 5 test questions from the set of 10 possible problems is $\binom{10}{5}$, each of which is equally likely. Similarly, the number of ways to choose 5 test questions from the set of 7 which the student knows how to solve is $\binom{7}{5}$. Therefore, the probability that a randomly selected combination of 5 test problems equals some combination of 5 of the 7 questions the student knows is

$$\frac{\binom{7}{5}}{\binom{10}{5}} = \frac{\frac{7!!}{2!5!}}{\frac{10!}{5!5!}} = \frac{7!}{2!5!}\frac{5!5!}{10!} = \frac{7!}{2!}\frac{5!}{10!} = \frac{7\cdot 6\cdot 5\cdot 4\cdot 3}{10\cdot 9\cdot 8\cdot 7\cdot 6} = \frac{1}{12} \approx 8.33\%$$

   (b) Let C_5 = the event that the student answers all 5 problems correctly.
   Let C_4 = the event that the student answers exactly 4 of the 5 problems correctly.
   Let $C_{\geq 4}$ = the event that the student answers at least 4 of the 5 problems correctly.
   If the student answers exactly 4 of the 5 problems correctly, they cannot possibly have answered all 5 questions correctly, and vice-versa. Therefore, C_5 and C_4 are mutually disjoint.
   Also, if a student answers at least 4 of the 5 questions correctly, they could have either answered exactly 4 of the questions correctly or all 5 problems correctly, but not both. Thus, we know that $C_{\geq 4} = C_4 \cup C_5$.
   Since C_4 and C_5 are mutually disjoint, this implies the probability that a student answers at least 4 of the problems correctly is

$$\mathbb{P}(C_{\geq 4}) = \mathbb{P}(C_4) + \mathbb{P}(C_5)$$

   From part (a), we already know that

$$\mathbb{P}(C_5) = \frac{1}{12}$$

   so we just need to calculate $\mathbb{P}(C_4)$. We can do this similarly to the alternative approach from the end of part (a).
   Once again, there are $\binom{10}{5}$ ways to choose 5 test problems from the set of 10 possible problems. However, this time, we want to choose only 4 problems from the set of 7 problems which the student knows how

to solve, which can be done in $\binom{7}{4}$ ways. For each of these $\binom{7}{4}$ combinations of 4 problems which the student knows how to solve, we need to choose one problem from the set of $10-7=3$ problems which the student cannot solve, which can be done in $\binom{3}{1}=3$ ways. Thus, the total number of ways that the student can get exactly 4 of the 5 problems right is $|C_4| = 3\binom{7}{4}$. Since all of the $\binom{10}{5}$ combinations of test questions are equally likely, this implies the probability that the student gets exactly 4 out of the 5 questions correct is

$$\mathbb{P}(C_4) = \frac{3\binom{7}{4}}{\binom{10}{5}} = \frac{\frac{3*7!}{4!3!}}{\frac{10!}{5!5!}} = \frac{7*6*5}{2}\frac{5!5!}{10!} = \frac{7*3*5*5!}{10*9*8*7*6} = \frac{5!}{2*3*8*6} = \frac{5}{2*6} = \frac{5}{12}$$

Combining this with the result from part (a), we find that the probability that the student gets at least 4 out of the 5 problems correct is

$$\mathbb{P}(C_{\geq 4}) = \mathbb{P}(C_4) + \mathbb{P}(C_5) = \frac{1}{12} + \frac{5}{12} = \frac{6}{12} = \frac{1}{2} = 50\%$$

5. (Ross P2.43)

   (a) If N people, including A and B, are randomly arranged in a line, what is the probability that A and B are next to each other?

   (b) What would the probability be if the people were randomly arranged in a circle?

Solution.
(a) If N people are randomly arranged in a line, then there are $N!$ total permutations of the people, each of which is equally likely. Therefore, the size of our sample space, S, is $|S| = N!$.
Now, we need to count the number of permutations in which A and B are next to each other. Let C = the event that A and B are next to each other when the N people are randomly arranged in a line. There are $N-1$ pairs of adjacent people. If we let the first person in line be p_1, the second be p_2, and so on, all the way until p_N, then these $N-1$ pairs of adjacent people are

$$(p_1, p_2), (p_2, p_3), ..., (p_{N-1}, p_N)$$

We need to choose one of these $N-1$ pairs to be the locations of A and B, which can be done in $\binom{N-1}{1} = N-1$ ways. For each of these choices, we have $2! = 2$ choices for the permutation of A and B within the pair. Once we make this choice, we have to permute the remaining $N-2$ people in the remaining $N-2$ places in the line. This can be done in $(N-2)!$ ways. Therefore, the total number of permutations in which A and B are next to each other when the N people are randomly arranged in a line is

$$|C| = (N-1) \cdot 2 \cdot (N-2)! = 2(N-1)!$$

Since all permutations are equally likely, this implies that the probability that A and B are next to each other when the N people are randomly arranged in a line is

$$\mathbb{P}(C) = \frac{|C|}{|S|} = \frac{2(N-1)!}{N!} = \frac{2}{N}$$

(b) Let's number the N spots in the circle from 1 to N. Then we still have $N!$ total ways to permute the N people among the N spots in the circle, each of which is equally likely. Therefore, the size of our sample space, S', is $|S'| = N!$.
Now, we need to count the number of permutations in which A and B are next to each other, considering that the people are now arranged in a circle. Let D = the event that A and B are next to each other when the N people are randomly arranged in a circle. Now, there are N pairs of adjacent people. If we let the person in spot 1 be p_1, the person in spot 2 be p_2, and so on, until we let the person in spot N be p_N, then the N pairs of adjacent people are

$$(p_1, p_2), (p_2, p_3), ..., (p_{N-1}, p_N), (p_N, p_1)$$

We need to choose one of these N pairs to be the locations of A and B, which can be done in $\binom{N}{1} = N$ ways. For each of these choices, we have $2! = 2$ choices for the permutation of A and B within the pair. Once we make this choice, we have to permute the remaining $N-2$ people in the remaining $N-2$ spots in the circle. This can be done in $(N-2)!$ ways. Therefore, the total number of permutations in which A and B are next to each other when the N people are randomly arranged in a circle is

$$|D| = N \cdot 2 \cdot (N-2)!$$

Since all permutations are equally likely, this implies that the probability that A and B are next to each other when the N people are randomly arranged in a circle is

$$\mathbb{P}(D) = \frac{|D|}{|S'|} = \frac{N \cdot 2 \cdot (N-2)!}{N!} = \frac{2}{N-1}$$

6. (Ross P2.45) A woman has n keys, of which one will open her door.

   (a) If she tries the keys at random, discarding those that do not work, what is the probability that she will open the door on her kth try?

   (b) What if she does not discard previously tried keys?

Solution.
(a) The woman is guaranteed to open the door by the nth try, so the probability of her opening the door on her kth try only makes sense for $1 \leq k \leq n$. For such k, if the woman opens the door on the kth try, we know that she did not open the door on any of her first $k-1$ tries. On the first try, there is a $\frac{n-1}{n}$ probability that she does not open the door. After discarding the first key, there is an $\frac{n-2}{n-1}$ probability that she does not open the door on her second try. After discarding the second key, there is an $\frac{n-3}{n-2}$ probability that she does not open the door on her third try. This pattern continues until there is a $\frac{n-k+1}{n-k+2}$ probability that she doesn't open the door on her $k-1$th try. She must open the door on her kth try, and there are $n-k+1$ keys left to choose from, only one of which will open the door. Thus, the probability that she opens the door on her kth try, given that she already discarded the keys from her first $k-1$ failed attempts, is $\frac{1}{n-k+1}$. Multiplying these probabilities together, we find that the total probability that the woman opens the door on her kth try is

$$\frac{n-1}{n}\frac{n-2}{n-1}\frac{n-3}{n-2}\cdots\frac{n-k+1}{n-k+2}\frac{1}{n-k+1} = \frac{1}{n}$$

Alternatively, we could note that there are $\frac{n!}{(n-k)!}$ permutations of k keys from a set of n keys. To form a permutation such that the woman opens the door on the kth try, the woman must try $k-1$ keys from the $n-1$ keys that *do not* open the door. There are $\frac{(n-1)!}{((n-1)-(k-1))!} = \frac{(n-1)!}{(n-k)!}$ ways to do this. Since the woman must choose the key that opens the door on her kth try, there is only one option for the first try. This leaves a total of $\frac{(n-1)!}{(n-k)!}$ ways to permute k of the n keys such that the woman opens the door on her kth try. Since all of the permutations are equally likely, the probability that the woman opens the door on her kth try is

$$\frac{\frac{(n-1)!}{(n-k)!}}{\frac{n!}{(n-k)!}} = \frac{(n-1)!}{(n-k)!}\frac{(n-k)!}{n!} = \frac{(n-1)!}{n!} = \frac{1}{n}$$

The two methods produce the same probability, as expected.

(b) If the woman doesn't discard previously tried keys, then the probability that the woman opens the door on any given roll is $\frac{1}{n}$, and the probability that the woman doesn't open the door on any given roll is $\frac{n-1}{n}$. Since the woman opens the door on her kth roll, we know that she doesn't open the door on any of her first $k-1$ rolls. Thus, each of her first $k-1$ rolls has an independent probability of $\frac{n-1}{n}$, while her kth roll has an independent probability of $\frac{1}{n}$. Multiplying these probabilities together, we find that the probability that the woman opens the door on her kth try if she *does not* discard previously tried keys is

$$\left(\frac{n-1}{n}\right)^{k-1} \cdot \frac{1}{n}$$

7. (Ross P2.47) Suppose that 5 of the numbers $1, 2, \ldots, 14$ are chosen. Find the probability that 9 is the third smallest value chosen.

Solution. There are $\binom{14}{5}$ equally likely ways to choose 5 numbers from the 14 numbers in $\{1, 2, \ldots, 14\}$. Thus, the size of our sample space, S, is $|S| = \binom{14}{5}$.
Let E = the event that 9 is the third smallest value chosen. For any $e \in E$, 9 must be one of the numbers chosen, 2 numbers chosen must be smaller than 9, and 2 numbers chosen must be larger than 9. Since we know we have to choose 9, we only have to consider how many ways we can select the other 4 numbers. There are 8 numbers in $\{1, 2, \ldots, 14\}$ that are less than 9, from which we need to select 2 numbers. There are $\binom{8}{2}$ ways to do this. There are 5 numbers in $\{1, 2, \ldots, 14\}$ that are greater than 9, from which we need to select 2 numbers. There are $\binom{5}{2}$ ways to do this, and each of these choices is made independently from each of the $\binom{8}{2}$ choices for the 2 numbers smaller than 9. Therefore, the total number of combinations of 5 elements from $\{1, 2, \ldots, 14\}$ with 9 as the third smallest value in the combination is

$$|E| = \binom{8}{2}\binom{5}{2} = \frac{8!}{6!2!}\frac{5!}{2!3!} = 28 \cdot 10 = 280$$

Since all combinations of 5 elements from $\{1, 2, \ldots, 14\}$ are equally likely, this means the probability that a combination has 9 as its third smallest element is

$$\mathbb{P}(E) = \frac{|E|}{|S|} = \frac{280}{\binom{14}{5}} = \frac{280}{\frac{14!}{5!9!}} = 280 \cdot \frac{5!9!}{14!} = 280 \cdot \frac{5!}{14 \cdot 13 \cdot 12 \cdot 11 \cdot 10} = \frac{280}{14 \cdot 13 \cdot 11} = \frac{20}{143} \approx 13.99\%$$

8. Consider the claim, "The probability that a randomly selected integer is even is 1/2." Either rigorously justify the claim using the axioms of probability, or show that it cannot be done.

Solution.
This cannot be done using the axioms of probability. We will show this by applying the axioms of probability and deriving a contradiction.
The sample space S, is the set of all integers, so $S = \mathbb{Z}$.
By the second axiom of probability, we know that

$$\mathbb{P}(S) = 1$$

Let E_i = the event that the randomly selected integer is i, for all $i \in \mathbb{Z}$.
For all $i \in E_i$, $i \in \mathbb{Z} \implies i \in S$. Also, for all $i \in S$, $i \in \mathbb{Z} \implies i \in E_i$. Therefore, we know

$$S = \mathbb{Z} = \bigcup_{i=-\infty}^{\infty} E_i$$

In any outcome, if the randomly selected integer is i, then the randomly selected integer cannot be j for all $j \neq i$. Therefore, for all $i \in E_i$, $i \notin E_j$ for all $j \neq i$, and vice-versa. Thus, we know that the E_i's are mutually disjoint for all $i \in \mathbb{Z}$. By the third axiom of probability, we know

$$\mathbb{P}(S) = \mathbb{P}\Big(\bigcup_{i=-\infty}^{\infty} E_i\Big) = \sum_{i=-\infty}^{\infty} \mathbb{P}(E_i)$$

Since the integer is selected randomly, there is an equal probability of selecting each integer, for all $z \in \mathbb{Z}$. Therefore, the probabilities of all the E_i's should be equal. Thus, we know that

$$p = \cdots = \mathbb{P}(E_{-2}) = \mathbb{P}(E_{-1}) = \mathbb{P}(E_0) = \mathbb{P}(E_1) = \mathbb{P}(E_2) = \cdots$$

By the first axiom of probability, we know that

$$0 \leq p \leq 1$$

If $p = 0$, then

$$\mathbb{P}(S) = \sum_{i=-\infty}^{\infty} \mathbb{P}(E_i) = \sum_{i=-\infty}^{\infty} p = \sum_{i=-\infty}^{\infty} 0 = 0$$

which is a contradiction since the second axiom of probability guarantees that

$$\mathbb{P}(S) = 1$$

If $0 < p \leq 1$, then

$$\mathbb{P}(S) = \sum_{i=-\infty}^{\infty} \mathbb{P}(E_i) = \sum_{i=-\infty}^{\infty} p = \infty \cdot p = \infty$$

which is also a contradiction by the second axiom of probability.

Therefore, applying the axioms of probability yields a contradiction, so we cannot prove the claim using the axioms of probability.
Note: If we let E_e = the event that a randomly selected integer is even and try to calculate its probability, we similarly fail to prove $\mathbb{P}(E_e) = \frac{1}{2}$.
If $p = 0$, then

$$\mathbb{P}(E_e) = \sum_{i \in \mathbb{Z} \text{ s.t. } i \text{ is even}} p = \sum_{i \in \mathbb{Z} \text{ s.t. } i \text{ is even}} 0 = 0 \neq \frac{1}{2}$$

If $p > 0$, then

$$\mathbb{P}(E_e) = \sum_{i \in \mathbb{Z} \text{ s.t. } i \text{ is even}} p = \infty \cdot p = \infty \neq \frac{1}{2}$$

since there are infinitely many even integers.
Thus, it is impossible to prove that "The probability that a randomly selected integer is even is $\frac{1}{2}$" using the axioms of probability.

9. (Ross P2.55) Compute the probability that a hand of 13 cards contains

   (a) the ace and king of at least one suit;

   (b) all 4 of at least 1 of the 13 denominations.

Solution.
(a) There are $\binom{52}{13}$ equally likely ways to choose a hand of 13 cards from a standard deck of 52 cards. Therefore, the size of our sample space, S, is $|S| = \binom{52}{13}$.
Let E = the event that a randomly selected hand contains the ace and king of at least one suit. We want to find $|E|$.
Let E_H = the event that a randomly selected hand contains the ace and king of hearts.
Let E_D = the event that a randomly selected hand contains the ace and king of diamonds.
Let E_C = the event that a randomly selected hand contains the ace and king of clubs.
Let E_S = the event that a randomly selected hand contains the ace and king of hearts.
For any $e \in E$, e could contain exactly one suited ace king pair, or it could contain two of the suited ace king pairs, or it could contain three of the suited ace king pairs, or it could contain all four of the suited ace king pairs. Therefore, we know that the event that a randomly selected hand contains at least one suited ace king pair is

$$E = E_H \cup E_D \cup E_C \cup E_S$$

Since an individual hand could have more than one suited ace king pair, E_H, E_D, E_C, and E_S are not mutually disjoint. Therefore, to calculate $\mathbb{P}(E) = \mathbb{P}(E_H \cup E_D \cup E_C \cup E_S)$, we need to apply the Principle of Inclusion-Exclusion. By the Principle of Inclusion-Exclusion, we know that

$$\begin{aligned}
\mathbb{P}(E) = \quad & \mathbb{P}(E_H \cup E_D \cup E_C \cup E_S) = \sum_{k=1}^{4} \left((-1)^{k-1} \sum_{i_1,\ldots,i_k \in \{H,D,C,S\}} \mathbb{P}(E_{i_1}) \cap \cdots \cap \mathbb{P}(E_{i_k}) \right) \\
= \quad & \mathbb{P}(E_H) + \mathbb{P}(E_D) + \mathbb{P}(E_C) + \mathbb{P}(E_S) - \mathbb{P}(E_H \cap E_D) - \mathbb{P}(E_H \cap E_C) - \mathbb{P}(E_H \cap E_S) \\
& - \mathbb{P}(E_D \cap E_C) - \mathbb{P}(E_D \cap E_S) - \mathbb{P}(E_C \cap E_S) + \mathbb{P}(E_H \cap E_D \cap E_C) + \mathbb{P}(E_H \cap E_D \cap E_S) \\
& + \mathbb{P}(E_H \cap E_C \cap E_S) + \mathbb{P}(E_D \cap E_C \cap E_S) - \mathbb{P}(E_H \cap E_D \cap E_C \cap E_S) \qquad (1)
\end{aligned}$$

We can now calculate each of these probabilities individually:
For any $e \in E_H$, we must pick 2 of the 13 cards to be the ace and king of hearts, which can be done in 1 way. We can choose the remaining 11 cards from the remaining 50 cards in the deck, which can be done in $\binom{50}{11}$ ways. Thus, there are a total of $\binom{50}{11}$ 13 card hands that contain the ace and king of hearts, so $|E_H| = \binom{50}{11}$. Since all hands are equally likely, this means that

$$\mathbb{P}(E_H) = \frac{|E_H|}{|S|} = \frac{\binom{50}{11}}{\binom{52}{13}}$$

By the symmetry of a standard deck of playing cards, we also know

$$\mathbb{P}(E_H) = \mathbb{P}(E_D) = \mathbb{P}(E_C) = \mathbb{P}(E_S)$$

Therefore, we find that

$$\mathbb{P}(E_H) + \mathbb{P}(E_D) + \mathbb{P}(E_C) + \mathbb{P}(E_S) = 4\mathbb{P}(E_H) = 4\frac{\binom{50}{11}}{\binom{52}{13}}$$

For any $e \in (E_H \cap E_D)$, we must pick 4 of the 13 cards to be the ace and king of hearts and the ace and king of diamonds, which can be done in 1 way. We can choose the remaining 9 cards in our hand from the remaining 48 cards in the deck, which can be done in $\binom{48}{9}$ ways. Thus, there are a total of $\binom{48}{9}$ 13 card hands that contain the ace and king of hearts and the ace and king of diamonds, so $|E_H \cap E_D| = \binom{48}{9}$. Since all hands are equally likely, this implies that

$$\mathbb{P}(E_H \cap E_D) = \frac{|E_H \cap E_D|}{|S|} = \frac{\binom{48}{9}}{\binom{52}{13}}$$

By the symmetry of a standard deck of playing cards, we know

$$\mathbb{P}(E_H \cap E_D) = \mathbb{P}(E_H \cap E_C) = \mathbb{P}(E_H \cap E_S) = \mathbb{P}(E_D \cap E_C) = \mathbb{P}(E_D \cap E_S) = \mathbb{P}(E_C \cap E_S)$$

Therefore, we find that

$$\begin{aligned} & -\mathbb{P}(E_H \cap E_D) - \mathbb{P}(E_H \cap E_C) - \mathbb{P}(E_H \cap E_S) - \mathbb{P}(E_D \cap E_C) - \mathbb{P}(E_D \cap E_S) - \mathbb{P}(E_C \cap E_S) \\ = \quad & -6\mathbb{P}(E_H \cap E_D) = -6\frac{\binom{48}{9}}{\binom{52}{13}} \end{aligned}$$

For any $e \in (E_H \cap E_D \cap E_C)$, we must pick 6 of the 13 cards to be the ace and king of hearts, the ace and king of diamonds, and the ace and king of clubs, which can be done in 1 way. We can choose the remaining 7 cards in our hand from the remaining 46 cards in the deck, which can be done in $\binom{46}{7}$ ways. Thus, there are a total of $\binom{46}{7}$ 13 card hands that contain the ace and king of hearts, the ace and king of diamonds, and the ace and king of clubs, so $|E_H \cap E_D \cap E_C| = \binom{46}{7}$. Since all hands are equally likely, this implies that

$$\mathbb{P}(E_H \cap E_D \cap E_C) = \frac{|E_H \cap E_D \cap E_C|}{|S|} = \frac{\binom{46}{7}}{\binom{52}{13}}$$

By the symmetry of a standard deck of playing cards, we know

$$\mathbb{P}(E_H \cap E_D \cap E_C) = \mathbb{P}(E_H \cap E_D \cap E_S) = \mathbb{P}(E_H \cap E_C \cap E_S) = \mathbb{P}(E_D \cap E_C \cap E_S)$$

Therefore, we find that

$$\begin{aligned} & \mathbb{P}(E_H \cap E_D \cap E_C) + \mathbb{P}(E_H \cap E_D \cap E_S) + \mathbb{P}(E_H \cap E_C \cap E_S) + \mathbb{P}(E_D \cap E_C \cap E_S) \\ = \quad & 4\mathbb{P}(E_H \cap E_D \cap E_C) = 4\frac{\binom{46}{7}}{\binom{52}{13}} \end{aligned}$$

For any $e \in (E_H \cap E_D \cap E_C \cap E_S)$, we must pick 8 of the 13 cards to be the ace and king of hearts, the ace and king of diamonds, the ace and king of clubs, and the ace and king of spades, which can be done in 1 way. We can choose the remaining 5 cards in our hand from the remaining 44 cards in the deck, which can be done in $\binom{44}{5}$ ways. Thus, there are a total of $\binom{44}{5}$ 13 card hands that contain the ace and king of hearts, the ace and king of diamonds, the ace and king of clubs, and the ace and king of spades, so $|E_H \cap E_D \cap E_C \cap E_S| = \binom{44}{5}$. Therefore,

$$\mathbb{P}(E_H \cap E_D \cap E_C \cap E_S) = \frac{|E_H \cap E_D \cap E_C \cap E_S|}{|S|} = \frac{\binom{44}{5}}{\binom{52}{13}}$$

Plugging these probabilities into (1), we find that the total probability that a randomly selected hand of 13 cards has at least one suited ace king pair is

$$\begin{aligned} \mathbb{P}(E) = \quad & \mathbb{P}(E_H \cup E_D \cup E_C \cup E_S) = 4\frac{\binom{50}{11}}{\binom{52}{13}} - 6\frac{\binom{48}{9}}{\binom{52}{13}} + 4\frac{\binom{46}{7}}{\binom{52}{13}} - \frac{\binom{44}{5}}{\binom{52}{13}} \\ = \quad & \frac{4\binom{50}{11} - 6\binom{48}{9} + 4\binom{46}{7} - \binom{44}{5}}{\binom{52}{13}} \approx 21.98\% \end{aligned}$$

(b) We will apply a similar argument as in part (a). Once again, our sample space S contains all possible 13 card hands for a total of $|S| = \binom{52}{13}$ equally likely hands.
For this problem, we will let 1 = ace, 11 = jack, 12 = queen, and 13 = king.
Let F = the event that a randomly selected 13 card hand contains all 4 of at least 1 of the 13 denominations. Let F_i = the event that a randomly selected 13 card hand contains all 4 cards of denomination

i, for all $1 \leq i \leq 13$. Then the event that a randomly selected 13 card hand contains all 4 of at least 1 of the 13 denominations is

$$F = \bigcup_{i=1}^{13} F_i$$

Since the F_i's are not mutually disjoint, we apply the Principle of Inclusion-Exclusion to find

$$\mathbb{P}(F) = \mathbb{P}(\bigcup_{i=1}^{13} F_i) = \sum_{k=1}^{13} \left((-1)^{k-1} \sum_{1 \leq i_1 \leq \ldots \leq i_k \leq 13} \mathbb{P}(F_{i_1} \cap \ldots \cap F_{i_k}) \right)$$

Note: In order to have all 4 cards of 4 distinct denominations, we would need to have $4 \cdot 4 = 16$ cards in our hand. Since our hands only has 13 cards, the maximum number of distinct denominations that any hand can contain all 4 of is 3. Thus, $\mathbb{P}(F_{i_1} \cap \cdots \cap F_{i_k}) = 0$ for all $k \geq 4$. This implies that

$$\begin{aligned} \mathbb{P}(F) = \quad & \mathbb{P}(\bigcup_{i=1}^{13} F_i) = \sum_{k=1}^{3} \left((-1)^{k-1} \sum_{1 \leq i_1 \leq \ldots \leq i_k \leq 13} \mathbb{P}(F_{i_1} \cap \ldots \cap F_{i_k}) \right) \\ = \quad & \sum_{i=1}^{13} \mathbb{P}(F_i) - \sum_{1 \leq i_1 \leq i_2 \leq 13} \mathbb{P}(F_{i_1} \cap F_{i_2}) + \sum_{1 \leq i_1 \leq i_2 \leq i_3 \leq 13} \mathbb{P}(F_{i_1} \cap F_{i_2} \cap F_{i_3}) \qquad (1) \end{aligned}$$

Now, let's calculate these probabilities individually.
For any $f \in F_i$, we must pick 4 out of our 13 cards to be the 4 cards of denomination i, which can be done in one way. We can then choose our remaining 9 cards from the remaining 48 cards in the deck, which can be done in $\binom{48}{9}$ equally likely ways. Thus, the total number of ways that a randomly selected hand of 13 cards contains all 4 cards of denomination i is $|F_i| = \binom{48}{9}$ for all $1 \leq i \leq 13$. Since each hand is equally likely, this implies the probability that a randomly selected 13 card hand contains all 4 cards of denomination i is

$$\mathbb{P}(F_i) = \frac{|F_i|}{|S|} = \frac{\binom{48}{9}}{\binom{52}{13}}$$

for all $1 \leq i \leq 13$. Therefore, we know that

$$\sum_{i=1}^{13} \mathbb{P}(F_i) = 13\mathbb{P}(F_i) = 13\frac{\binom{48}{9}}{\binom{52}{13}}$$

For any $f \in (F_i \cap F_j)$, where $i \neq j$, we must pick 8 out of our 13 cards to be the 4 cards of denomination i and the 4 cards of denomination j, which can be done in one way. We then need to choose our remaining 5 cards from the remaining 44 cards in the deck, which can be done in $\binom{44}{5}$ equally likely ways. Thus, the total number of ways that a randomly selected hand of 13 cards contains all 4 cards of denomination i and all 4 cards of denomination j is $|F_i \cap F_j| = \binom{44}{5}$ for all $1 \leq i, j \leq 13$ s.t. $i \neq j$. Since each hand is equally likely, this implies that the probability that a randomly selected 13 card hand contains all 4 cards of denomination i and all 4 cards of denomination j is

$$\mathbb{P}(F_i \cap F_j) = \frac{|F_i \cap F_j|}{|S|} = \frac{\binom{44}{5}}{\binom{52}{13}}$$

This is true for all $1 \leq i, j \leq 13$ s.t. $i \neq j$ due to the symmetry of a standard deck of cards. There are $\binom{13}{2}$ ways to choose which 2 denominations are guaranteed to have all 4 of their cards contained in our hand. Thus, we know

$$\sum_{1 \leq i_1 \leq i_2 \leq 13} \mathbb{P}(F_{i_1} \cap F_{i_2}) = \binom{13}{2} \frac{\binom{44}{5}}{\binom{52}{13}}$$

For all $f \in (F_i \cap F_j \cap F_k)$, where $i \neq j \neq k$, we must pick 12 out of the 13 cards to be the 4 cards of denomination i, the 4 cards of denomination j, and the 4 cards of denomination k, which can be done

in one way. We can then choose our remaining 1 card from the remaining 40 cards in the deck, which can be done in $\binom{40}{1} = 40$ ways. Therefore, the total number of ways that a randomly selected hand of 13 cards contains all 4 cards of denomination i, j, and k is $|F_i \cap F_j \cap F_k| = 40$ for all $1 \leq i, j, k \leq 13$ s.t. $i \neq j \neq k$. Since each hand is equally likely, we know

$$\mathbb{P}(F_i \cap F_j \cap F_k) = \frac{|F_i \cap F_j \cap F_k|}{|S|} = \frac{40}{\binom{52}{13}}$$

This is true for all $1 \leq i, j, k \leq 13$ s.t. $i \neq j \neq k$ due to the symmetry of a standard deck of cards. There are $\binom{13}{3}$ ways to select which 3 of the denominations are guaranteed to have all 4 of their cards contained in our hand. Thus, we know

$$\sum_{1 \leq i_1 \leq i_2 \leq i_3 \leq 13} \mathbb{P}(F_{i_1} \cap F_{i_2} \cap F_{i_3}) = \binom{13}{3} \frac{40}{\binom{52}{13}}$$

Plugging these probabilities into (1), we find that the total probability that a randomly selected 13 card hand contains all 4 cards of at least 1 out of the 13 denominations is

$$\mathbb{P}(F) = \mathbb{P}(\bigcup_{i=1}^{13} F_i) = 13\frac{\binom{48}{9}}{\binom{52}{13}} - \binom{13}{2}\frac{\binom{44}{5}}{\binom{52}{13}} + \binom{13}{3}\frac{40}{\binom{52}{13}} = \frac{13\binom{48}{9} - \binom{13}{2}\binom{44}{5} + \binom{13}{3} \cdot 40}{\binom{52}{13}} \approx 3.42\%$$

10. You are playing a game with your younger brother which involves repeatedly rolling a 4-sided die. At the start of the game, he chose 3 and you chose 1. Every turn, the die is rolled; if it is a 3 or a 1, you or your brother gets to move, and otherwise a dragon can attack.

    Your brother is doing amazingly well—he's gotten many moves, and there have been few dragon attacks. Perhaps he's doing too well—you begin to suspect he is cheating with a special die! At the end of the game, the frequency of each number rolled is as follows:

Roll	Frequency
1	58
2	64
3	88
4	40

    Is this strong evidence that your brother is cheating? Justify your answer.

    Solution.
    In the case where my brother isn't cheating, each of the 4 values on the die should be equally likely on each roll. Therefore, after n rolls, we would expect $\frac{n}{4}$ ones, $\frac{n}{4}$ twos, $\frac{n}{4}$ threes, and $\frac{n}{4}$ fours.
    In our sample game, there are $58 + 64 + 88 + 40 = 250$ total rolls of the die. Thus we would expect the frequency of each possible value to be $\approx \frac{250}{4} = 62.5$. To determine if we have strong evidence that our brother is cheating, we should examine the value whose frequency differs the most from this expectation. Of any of the individual frequencies, this frequency should provide the strongest evidence for cheating. The differences between the expected number of rolls and the sample frequencies are:

    (i) Roll value $= 1$: $|58 - 62.5| = 4.5$
    (ii) Roll value $= 2$: $|64 - 62.5| = 1.5$
    (iii) Roll value $= 3$: $|88 - 62.5| = 25.5$
    (iv) Roll value $= 5$: $|40 - 62.5| = 22.5$

    Thus, frequency that differs most from the expected frequency is the frequency of the value 3, which occurs 25.5 times *more* than expected. To determine if this provides strong evidence that our brother is cheating, we want to find the probability that a truly random sequence of 250 fair 4-sided dice rolls results in at least 88 3's. On any given roll, the probability of rolling a 3 is $\frac{1}{4}$ and the probability of not rolling a 3 is $\frac{3}{4}$. If exactly k 3's are rolled, then there are $\binom{250}{k}$ choices for which specific dice rolls have value 3. Thus, the probability that a truly random sequence of 250 fair 4-sided dice rolls results in exactly k 3's is

$$\binom{250}{k}\left(\frac{1}{4}\right)^k\left(\frac{3}{4}\right)^{250-k}$$

    Therefore, the probability that a truly random sequence of 250 fair 4-sided dice rolls results in at least k 3's is

$$\sum_{i=k}^{250}\binom{250}{i}\left(\frac{1}{4}\right)^i\left(\frac{3}{4}\right)^{250-i}$$

    so the probability that a truly random sequence of 250 fair 4-sided dice rolls results in at least 88 3's is

$$\sum_{i=88}^{250}\binom{250}{i}\left(\frac{1}{4}\right)^i\left(\frac{3}{4}\right)^{250-i} \approx 0.02\%$$

    In lecture, we discussed the convention that a result with k successes over n trials is *suspiciously high* if the probability of at least k successes over n trials is $\leq 5\%$. Comparing this cutoff with the probability that a truly random sequence of 250 fair 4-sided dice rolls results in at least 88 3's, we clearly see that

$$0.02\% < 5\%$$

indicates that number of 3's rolled in our game is indeed *suspiciously high*. In fact, since the probability of rolling at least 88 3's is $\approx 0.02\%$, there is only a $\approx 0.02\%$ chance of obtaining a number of 3's as high as what my brother got in our sample game. Since getting more 3's directly helps my brother in the game, and there is such a small chance of rolling as many 3's as my brother rolled by pure chance, we do have strong evidence that my brother is cheating.

Assignment 6

Math 407 (Swanson) – Spring 2023
Homework 1
Due Friday 1/13, 11:59pm

Name: Emerson Kahle Section: 39981

- You must upload your solutions to Gradescope as **one single, high-quality PDF**. You can convert paper-based work to a high-quality PDF using a scanning app for mobile devices, such as Adobe Scan (free, available for iOS and Android, can do multiple pages) or many others. If necessary, you can combine or merge multiple PDF's into a single PDF using a variety of services, such as Adobe Acrobat's cloud-based merge tool.
- After you upload, you must match each question with its corresponding page using Gradescope's interface. This allows graders to spend more time giving you feedback instead of hunting through submissions.
- Answers without supporting work will receive no credit. Show your work.
- You are encouraged to work together on homework, but **you must write up your solutions separately in your own words**. Copying from your fellow students or other sources is a serious academic integrity violation. In particular, you may not use "tutoring" services which simply provide answers.
- You are encouraged to typeset your solutions in LaTeX. Source code has been provided on Blackboard. Overleaf is a popular cloud-based editor.
- Problem numbers refer to the course textbook, though the problems may have been modified significantly.

1. (Ross P3.1 and P3.2)

   (a) Two fair dice are rolled. What is the conditional probability that at least one lands on 6 given that the dice land on different numbers?

   (b) If two fair dice are rolled, what is the conditional probability that the first one lands on 6 given that the sum of the dice is i? Compute for all values of i between 2 and 12.

   (c) Now compute the conditional probability that the sum of the dice is i given that the first one lands on 6, for each i from 2 to 12. What is the sum of the resulting answers? Relate this to the Law of Total Probability.

   Solution.
   (a) Let $E =$ the event that at least one of the two dice lands on 6.
   Let $F =$ the event that the two dice land on different numbers.
   Then the conditional probability that at least one lands on 6 given that the dice land on different numbers is

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(EF)}{\mathbb{P}(F)} \qquad (1)$$

   The sample space for a pair of dice rolls is

$$\begin{aligned} S = \{1,2,3,4,5,6\}^2 = \quad & \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\ & (2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\ & (3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\ & (4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\ & (5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\ & (6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\} \end{aligned}$$

   where each outcome is equally likely since the dice are fair.
   We can clearly see that the set of outcomes where the two dice land on different numbers is

$$\begin{aligned} F = S - \cup_{i=1}^{6}(i,i) = \quad & \{(1,2),(1,3),(1,4),(1,5),(1,6),\\ & (2,1),(2,3),(2,4),(2,5),(2,6),\\ & (3,1),(3,2),(3,4),(3,5),(3,6),\\ & (4,1),(4,2),(4,3),(4,5),(4,6),\\ & (5,1),(5,2),(5,3),(5,4),(5,6),\\ & (6,1),(6,2),(6,3),(6,4),(6,5)\} \end{aligned}$$

   so $|F| = 30$. Since all outcomes are equally likely, we know the probability that the dice land on different numbers is

$$\mathbb{P}(F) = \frac{|F|}{|S|} = \frac{30}{36} = \frac{5}{6} \approx 83.33\% \qquad (2)$$

   Also, we can see that the set of outcomes where at least one die lands on 6 and the two dice land on different numbers is

$$\begin{aligned} EF = \bigcup_{1\le i\ne j\le 6}(i,j) = \quad & \{(1,6),(2,6),(3,6),(4,6),(5,6),\\ & (6,5),(6,4),(6,3),(6,2),(6,1)\} \end{aligned}$$

   so $|EF| = 10$. Since all outcomes are equally likely, we know the probability that the at least one die lands on 6 and the dice land on different numbers is

$$\mathbb{P}(EF) = \frac{|EF|}{|S|} = \frac{10}{36} = \frac{5}{18} \approx 27.78\% \qquad (3)$$

Plugging (2) and (3) into (1), we find the conditional probability that at least one lands on 6 given that the dice land on different numbers is

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(EF)}{\mathbb{P}(F)} = \frac{\frac{10}{36}}{\frac{30}{36}} = \frac{10}{36}\frac{36}{30} = \frac{10}{30} = \frac{1}{3} \approx 33.33\%$$

(b) Let E_i = the event that the sum of the dice is i for all $2 \leq i \leq 12$.
Let F = the event that the first die lands on 6.
Then the conditional probability that the first one lands on 6 given that the sum of the dice is i is

$$\mathbb{P}(F|E_i) = \frac{\mathbb{P}(FE_i)}{\mathbb{P}(E_i)} \qquad (4)$$

Let's calculate this for each $2 \leq i \leq 12$.
The sample space is the same as in part (a), and all outcomes are still equally likely.
There are only 6 ways that the first die can land on 6, so

$$F = \bigcup_{i=1}^{6}(6, i) = \{(6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}$$

There is only one way to have a sum of 2, so

$$E_2 = \{(1,1)\}, \quad |E_2| = 1$$

Therefore, $\mathbb{P}(E_2) = \frac{|E_2|}{|S|} = \frac{1}{36} \approx 2.78\%$.
Clearly, for all $e \in E_2$, $e \notin F$, so $FE_2 = \emptyset$, so $\mathbb{P}(FE_2) = 0$.
Plugging this into (4), we see that

$$\mathbb{P}(F|E_2) = \frac{0}{\frac{1}{36}} = 0 = 0\%$$

There are two ways to have a sum of 3, so

$$E_3 = \{(2,1), (1,2)\} \quad |E_3| = 2$$

Therefore, $\mathbb{P}(E_3) = \frac{|E_3|}{|S|} = \frac{2}{36} = \frac{1}{18} \approx 5.56\%$. Clearly, for all $e \in E_3$, $e \notin F$, so $FE_3 = \emptyset$, so $\mathbb{P}(FE_3) = 0$.
Plugging this into (4), we see that

$$\mathbb{P}(F|E_3) = \frac{0}{\frac{1}{18}} = 0 = 0\%$$

There are three ways to have a sum of 4, so

$$E_4 = \{(3,1), (2,2), (1,3)\} \quad |E_4| = 3$$

Therefore, $\mathbb{P}(E_4) = \frac{|E_4|}{|S|} = \frac{3}{36} = \frac{1}{12} \approx 8.33\%$. Clearly, for all $e \in E_4$, $e \notin F$, so $FE_4 = \emptyset$, so $\mathbb{P}(FE_4) = 0$.
Plugging this into (4), we see that

$$\mathbb{P}(F|E_4) = \frac{0}{\frac{1}{12}} = 0 = 0\%$$

There are four ways to have a sum of 5, so

$$E_5 = \{(4,1), (3,2), (2,3)(1,4)\} \quad |E_5| = 4$$

Therefore, $\mathbb{P}(E_5) = \frac{|E_5|}{|S|} = \frac{4}{36} = \frac{1}{9} \approx 11.11\%$. Clearly, for all $e \in E_5$, $e \notin F$, so $FE_5 = \emptyset$, so $\mathbb{P}(FE_5) = 0$.
Plugging this into (4), we see that

$$\mathbb{P}(F|E_5) = \frac{0}{\frac{1}{9}} = 0 = 0\%$$

There are five ways to have a sum of 6, so

$$E_6 = \{(5,1),(4,2),(3,3),(2,4),(1,5)\} \quad |E_6| = 5$$

Therefore, $\mathbb{P}(E_6) = \frac{|E_6|}{|S|} = \frac{5}{36} \approx 13.89\%$. Clearly, for all $e \in E_6$, $e \notin F$, so $FE_6 = \emptyset$, so $\mathbb{P}(FE_6) = 0$. Plugging this into (4), we see that

$$\mathbb{P}(F|E_6) = \frac{0}{\frac{5}{36}} = 0 = 0\%$$

There are six ways to have a sum of 7, so

$$E_7 = \{(6,1),(5,2),(4,3),(3,4),(2,5),(1,6)\} \quad |E_7| = 6$$

Therefore, $\mathbb{P}(E_7) = \frac{|E_7|}{|S|} = \frac{6}{36} = \frac{1}{6} \approx 16.67\%$. Clearly, $(6,1)$ is the only outcome in both E_7 and F, so $FE_7 = \{(6,1)\}$, so

$$\mathbb{P}(FE_7) = \frac{|FE_7|}{|S|} = \frac{1}{36} \approx 2.78\%$$

Plugging $\mathbb{P}(E_7)$ and $\mathbb{P}(FE_7)$ in (4), we see that

$$\mathbb{P}(F|E_7) = \frac{\frac{1}{36}}{\frac{6}{36}} = \frac{1}{36}\frac{36}{6} = \frac{1}{6} \approx 16.67\%$$

There are five ways to have a sum of 8, so

$$E_8 = \{(6,2),(5,3),(4,4),(3,5),(2,6)\} \quad |E_8| = 5$$

Therefore, $\mathbb{P}(E_8) = \frac{|E_8|}{|S|} = \frac{5}{36} \approx 13.89\%$. Clearly, $(6,2)$ is the only outcome in both E_8 and F, so $FE_8 = \{(6,2)\}$, so

$$\mathbb{P}(FE_8) = \frac{|FE_8|}{|S|} = \frac{1}{36} \approx 2.78\%$$

Plugging $\mathbb{P}(E_8)$ and $\mathbb{P}(FE_8)$ in (4), we see that

$$\mathbb{P}(F|E_8) = \frac{\frac{1}{36}}{\frac{5}{36}} = \frac{1}{36}\frac{36}{5} = \frac{1}{5} = 20\%$$

There are four ways to have a sum of 9, so

$$E_9 = \{(6,3),(5,4),(4,5),(3,6)\} \quad |E_9| = 4$$

Therefore, $\mathbb{P}(E_9) = \frac{|E_9|}{|S|} = \frac{4}{36} = \frac{1}{9} \approx 11.11\%$. Clearly, $(6,3)$ is the only outcome in both E_9 and F, so $FE_9 = \{(6,3)\}$, so

$$\mathbb{P}(FE_9) = \frac{|FE_9|}{|S|} = \frac{1}{36} \approx 2.78\%$$

Plugging $\mathbb{P}(E_9)$ and $\mathbb{P}(FE_9)$ in (4), we see that

$$\mathbb{P}(F|E_9) = \frac{\frac{1}{36}}{\frac{4}{36}} = \frac{1}{36}\frac{36}{4} = \frac{1}{4} = 25\%$$

There are three ways to have a sum of 10, so

$$E_{10} = \{(6,4),(5,5),(4,6)\} \quad |E_{10}| = 3$$

Therefore, $\mathbb{P}(E_{10}) = \frac{|E_{10}|}{|S|} = \frac{3}{36} = \frac{1}{12} \approx 8.33\%$. Clearly, $(6,4)$ is the only outcome in both E_{10} and F, so $FE_{10} = \{(6,4)\}$, so

$$\mathbb{P}(FE_{10}) = \frac{|FE_{10}|}{|S|} = \frac{1}{36} \approx 2.78\%$$

Plugging $\mathbb{P}(E_{10})$ and $\mathbb{P}(FE_{10})$ in (4), we see that

$$\mathbb{P}(F|E_{10}) = \frac{\frac{1}{36}}{\frac{3}{36}} = \frac{1}{36}\frac{36}{3} = \frac{1}{3} \approx 33.33\%$$

There are two ways to have a sum of 11, so

$$E_{11} = \{(6,5),(5,6)\} \quad |E_{11}| = 2$$

Therefore, $\mathbb{P}(E_{11}) = \frac{|E_{11}|}{|S|} = \frac{2}{36} = \frac{1}{18} \approx 5.56\%$. Clearly, $(6,5)$ is the only outcome in both E_{11} and F, so $FE_{11} = \{(6,5)\}$, so

$$\mathbb{P}(FE_{11}) = \frac{|FE_{11}|}{|S|} = \frac{1}{36} \approx 2.78\%$$

Plugging $\mathbb{P}(E_{11})$ and $\mathbb{P}(FE_{11})$ in (4), we see that

$$\mathbb{P}(F|E_{11}) = \frac{\frac{1}{36}}{\frac{2}{36}} = \frac{1}{36}\frac{36}{3} = \frac{1}{2} = 50\%$$

There is one way to have a sum of 12, so

$$E_{12} = \{(6,6)\} \quad |E_{12}| = 1$$

Therefore, $\mathbb{P}(E_{12}) = \frac{|E_{12}|}{|S|} = \frac{1}{36} \approx 2.78\%$. Clearly, $(6,6)$ is the only outcome in both E_{12} and F, so $FE_{12} = \{(6,6)\}$, so

$$\mathbb{P}(FE_{12}) = \frac{|FE_{12}|}{|S|} = \frac{1}{36} \approx 2.78\%$$

Plugging $\mathbb{P}(E_{12})$ and $\mathbb{P}(FE_{12})$ in (4), we see that

$$\mathbb{P}(F|E_{12}) = \frac{\frac{1}{36}}{\frac{1}{36}} = \frac{1}{36}\frac{36}{1} = 1 = 100\%$$

(c) Using the same event and sample space definitions as in the part (b), we see that the conditional probability that the sum of the dice is i given that the first one lands on 6 is

$$\mathbb{P}(E_i|F) = \frac{\mathbb{P}(E_iF)}{\mathbb{P}(F)} = \frac{\mathbb{P}(FE_i)}{\mathbb{P}(F)} \qquad (5)$$

for all $2 \leq i \leq 12$. We already calculated the numerator for all $2 \leq i \leq 12$ in part (b). We also saw that $|F| = 6$ in part (b), and since all outcomes are equally likely, we know

$$\mathbb{P}(F) = \frac{|F|}{|S|} = \frac{6}{36} = \frac{1}{6} \approx 16.67\%$$

Now, we can directly calculate $\mathbb{P}(E_i|F)$ for all $2 \leq i \leq 12$ using (5).

(i)

$$\mathbb{P}(E_2|F) = \frac{\mathbb{P}(FE_2)}{\mathbb{P}(F)} = \frac{0}{\frac{1}{6}} = 0 = 0\%$$

(ii)
$$\mathbb{P}(E_3|F) = \frac{\mathbb{P}(FE_3)}{\mathbb{P}(F)} = \frac{0}{\frac{1}{6}} = 0 = 0\%$$

(iii)
$$\mathbb{P}(E_4|F) = \frac{\mathbb{P}(FE_4)}{\mathbb{P}(F)} = \frac{0}{\frac{1}{6}} = 0 = 0\%$$

(iv)
$$\mathbb{P}(E_5|F) = \frac{\mathbb{P}(FE_5)}{\mathbb{P}(F)} = \frac{0}{\frac{1}{6}} = 0 = 0\%$$

(v)
$$\mathbb{P}(E_6|F) = \frac{\mathbb{P}(FE_6)}{\mathbb{P}(F)} = \frac{0}{\frac{1}{6}} = 0 = 0\%$$

(vi)
$$\mathbb{P}(E_7|F) = \frac{\mathbb{P}(FE_7)}{\mathbb{P}(F)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6} \approx 16.67\%$$

(vii)
$$\mathbb{P}(E_8|F) = \frac{\mathbb{P}(FE_8)}{\mathbb{P}(F)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6} \approx 16.67\%$$

(viii)
$$\mathbb{P}(E_9|F) = \frac{\mathbb{P}(FE_9)}{\mathbb{P}(F)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6} \approx 16.67\%$$

(ix)
$$\mathbb{P}(E_{10}|F) = \frac{\mathbb{P}(FE_{10})}{\mathbb{P}(F)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6} \approx 16.67\%$$

(x)
$$\mathbb{P}(E_{11}|F) = \frac{\mathbb{P}(FE_{11})}{\mathbb{P}(F)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6} \approx 16.67\%$$

(xi)
$$\mathbb{P}(E_{12}|F) = \frac{\mathbb{P}(FE_{12})}{\mathbb{P}(F)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6} \approx 16.67\%$$

Adding these probabilities up, we find that

$$\sum_{i=2}^{12} \mathbb{P}(E_i|F) = 0+0+0+0+0+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6} = 6\frac{1}{6} = 1 = 100\% = \mathbb{P}(S) \quad (6)$$

This is directly related to the Law of Total Probability, which states that, for any event A, and any mutually disjoint events $B_1, B_2, ..., B_k$ s.t. $S = B_1 \cup B_2 \cup ... \cup B_k$,

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

This is essentially adding up the probabilities that an outcome o is in event A when $o \in B_i$ for any $1 \leq i \leq k$. Since $S = B_1 \cup B_2 \cup ... \cup B_k$, and the B_i's are mutually disjoint, this is just the probability that o is in A given o is in S, which is just the probability of A, as the law states.
In our case, we have the mutually disjoint events $E_2, ..., E_{12}$, and we also have $S = E_2 \cup ... \cup E_{12}$.

However, we are summing $\mathbb{P}(E_i|F)$ instead of $\mathbb{P}(A|B_i)$, and we do not multiply by $\mathbb{P}(F)$ like the Law of Total Probability does. If we did, we would get

$$\mathbb{P}(F)\sum_{i=2}^{12}\mathbb{P}(E_i|F) = \frac{1}{6} = \mathbb{P}(F)$$

which would be the probability that we get an outcome in $F(E_2 \cup ... \cup E_{12}) = FS = F$. If not, we are essentially restricting our point of view to the event F, then adding up the probabilities that any $o \in F$ is in some E_i for all $2 \leq i \leq 12$. This is just the probability that any $o \in F$ is in S, which is 100% as we found in (6). Essentially, this is just reversing the direction of the condition from the Law of Total Probability. Instead of considering the probability that an outcome is in A considering it is in S, which is $\mathbb{P}(A)$, we are considering the probability that an outcome is in S considering it is in F, which is 100%.

2. At the end of the classic American game show *Let's Make A Deal*, the host, Monty Hall, would offer a contestant a chance to win a large prize such as a new car. The prize was behind one of three doors, #1, #2, or #3, but goats were hiding behind the other two doors!

   Monty had the contestant initially choose one of the three doors. Then he would dramatically reveal a goat behind one of the two remaining doors. Finally he gave the contestant a choice: did they want to switch their guess to the final remaining door, or did they want to keep their original guess? The audience helpfully screamed "switch" or "stay" at the contestant while they deliberated. The contestant received whatever was behind the final door they chose, be it a goat or a car.

   Answering Monty's question, "'switch' or 'stay'?" has become known as the "Monty Hall Problem." So, should the contestant switch or stay put? Justify your answer probabilistically.

   Solution. The contestant should *switch*.
   We assume there is an equal probability of the prize being behind each door. Since there are three doors, this means there is a $\frac{1}{3}$ probability of the prize being behind each door. Let's call the door initially selected by the contestant d_i. Let D_i = the event that the prize is behind door d_i. By the previous line of reasoning, we know

$$\mathbb{P}(D_i) = \frac{1}{3} \implies \mathbb{P}(D_i^c) = 1 - \frac{1}{3} = \frac{2}{3}$$

   Since the prize must be behind one of the three doors, if it is not behind d_i, then it must be behind one of the other two doors. Therefore, D_i^c = the event that the prize is behind a door other than d_i.
   Note: Since Monty reveals a goat behind one of the two doors that aren't d_i, if the goat is behind one of these two doors, it will always be behind the last remaining door. Therefore, if the prize is behind one of the two doors that aren't d_i, then switching to the final remaining door will always cause the contestant to win.
   Therefore, if the contestant decides to *stay*, he will win if and only if the prize is behind d_i, so the probability that he wins is

$$\mathbb{P}(D_i) = \frac{1}{3}$$

   However, if the contestant decides to *switch*, then he will win if and only if the prize is behind one of the two doors other than d_i, so the probability he wins is

$$\mathbb{P}(D_i^c) = \frac{2}{3}$$

   We can clearly see that

$$\mathbb{P}(\text{contestant wins}|\text{stay}) = \mathbb{P}(D_i) = \frac{1}{3} < \frac{2}{3} = \mathbb{P}(D_i^c) = \mathbb{P}(\text{contestant wins}|\text{switch})$$

   So the contestant has a greater probability of winning the prize if he decides to *switch*, so he should decide to *switch*.

3. (Ross P3.14) Suppose that an ordinary deck of 52 cards (which contains 4 aces) is randomly divided into 4 hands of 13 cards each. We are interested in determining p, the probability that each hand has an ace. Let E_i be the event that the ith hand has exactly one ace. Determine $p = P(E_1E_2E_3E_4)$ by using the multiplication rule.

Solution.
Applying the multiplication rule, we find that

$$p = \mathbb{P}(E_1E_2E_3E_4) = \mathbb{P}(E_1)\mathbb{P}(E_2|E_1)\mathbb{P}(E_3|E_1E_2)\mathbb{P}(E_4|E_1E_2E_3) \qquad (1)$$

Let's calculate the 4 probabilities in the right-hand side individually.

(i) $\mathbb{P}(E_1)$: For the first, hand the sample space is all possible 13 card hands from a 52 card deck. Therefore, we know that the size of our sample space, S_1, is

$$|S_1| = \binom{52}{13}$$

where all outcomes are equally likely.
We must choose one of the 4 aces to be in the first hand. There are $\binom{4}{1}$ ways to do this. The remaining 12 cards in the hand must not include any more aces. Thus, we can only choose the 12 remaining cards from the 48 cards in the deck that aren't aces (2 through King of all suits). There are $\binom{48}{12}$ ways to choose these 12 cards. Thus, the total number of ways that the first hand can get exactly one ace is

$$\binom{4}{1}\binom{48}{12} = 4\binom{48}{12}$$

Since all outcomes are equally likely, we can calculate that the probability the first hand gets exactly one ace is

$$\mathbb{P}(E_1) = \frac{\binom{4}{1}\binom{48}{12}}{|S_1|} = \frac{4\binom{48}{12}}{\binom{52}{13}} \approx 43.88\%$$

(ii) $\mathbb{P}(E_2|E_1)$: There are only $52 - 13 = 39$ cards remaining in the deck from which the 13 cards in the second hand may be selected, and exactly 3 of them are aces. Therefore, we know that the size of our sample space, S_2, is

$$|S_2| = \binom{39}{13}$$

where all outcomes are equally likely.
We must choose one of the 3 remaining aces to be in this second hand. There are $\binom{3}{1}$ ways to do this. The remaining 12 cards in the hand can only include the remaining $39 - 3 = 36$ cards in the deck that aren't aces. There are $\binom{36}{12}$ ways to choose these 12 cards. Thus, the total number of ways that the second hand can get exactly one ace, given that the first hand got exactly one ace is

$$\binom{3}{1}\binom{36}{12} = 3\binom{36}{12}$$

Since all outcomes are equally likely, we can calculate that the probability that the second hand gets exactly one ace given that the first hand got exactly one ace, is

$$\mathbb{P}(E_2|E_1) = \frac{\binom{3}{1}\binom{36}{12}}{|S_2|} = \frac{3\binom{36}{12}}{\binom{39}{13}} \approx 46.23\%$$

(iii) $\mathbb{P}(E_3|E_1E_2)$: There are only $39 - 13 = 26$ cards remaining in the deck from which the 13 cards in the third hand may be selected, and exactly 2 of them are aces. Therefore, we know that the size of our sample space, S_3, is

$$|S_3| = \binom{26}{13}$$

where all outcomes are equally likely.
We must choose one of the 2 remaining aces to be in this third hand. There are $\binom{2}{1}$ ways to do this. The remaining 12 cards in the hand can only include the remaining $26-2=24$ cards in the deck that aren't aces. There are $\binom{24}{12}$ ways to choose these 12 cards. Thus, the total number of ways that the third hand can get exactly one ace, given that the first two hands got exactly one ace each, is

$$\binom{2}{1}\binom{24}{12} = 2\binom{24}{12}$$

Since all outcomes are equally likely, we can calculate that the probability that the third hand gets exactly one ace, given that the first two hands both got exactly one ace, is

$$\begin{aligned}\mathbb{P}(E_3|E_1E_2) = & \quad \frac{\binom{2}{1}\binom{24}{12}}{|S_3|} = \frac{2\binom{24}{12}}{\binom{26}{13}} = 2\frac{24(23)...(14)(13)}{12!}\frac{13!}{26(25)...(15)(14)} \\ = & \quad \frac{24(23)...(14)(13)}{(25)(24)...(15)(14)} = \frac{13}{25} = 52\%\end{aligned}$$

(iv) $\mathbb{P}(E_4|E_1E_2E_3)$: There are only 13 cards remaining in the deck from which the 13 cards in the fourth hand may be selected, exactly one of which is an ace. Thus, the fourth hand will just consist of all 13 of the remaining cards, so the fourth hand is guaranteed to have one ace. Thus, we know that the probability that the fourth hand has exactly one ace, given that the first three hands all have exactly one ace each, is

$$\mathbb{P}(E_4|E_1E_2E_3) = 1 = 100\%$$

Now, we can plug the probabilities calculated in (i), (ii), (iii), and (iv) into (1), we find that the probability that all 4 hands have exactly one ace is

$$\begin{aligned}p = & \quad \mathbb{P}(E_1E_2E_3E_4) = \mathbb{P}(E_1)\mathbb{P}(E_2|E_1)\mathbb{P}(E_3|E_1E_2)\mathbb{P}(E_4|E_1E_2E_3) \\ = & \quad \frac{4\binom{48}{12}}{\binom{52}{13}} \cdot \frac{3\binom{36}{12}}{\binom{39}{13}} \cdot \frac{13}{25} \cdot 1 \approx (0.4388) \cdot (0.4623) \cdot (0.52) \cdot (1) \approx 10.55\%\end{aligned}$$

4. While it is not possible to randomly sample from the whole infinite set of integers, it is possible to fix n and randomly sample from the finite set $[n] = \{1, 2, \ldots, n\}$. For instance, the probability that a randomly chosen integer from [5] is even is 2/5.

   Given a subset A of the integers, the *natural density* of A is defined to be

$$\lim_{n\to\infty} \frac{\#(A \cap [n])}{n},$$

   if the limit exists.

   (a) Interpret $\frac{\#(A\cap[n])}{n}$ probabilistically.

   (b) Compute the natural density of the set $2\mathbb{Z}_{>0} = \{2, 4, \ldots\}$ of positive even integers.

   (c) Compute the natural density of the set $\{1, 3, 5, \ldots\}$ of positive odd integers.

   (d) Compute the natural density of the set $\{1, 4, 9, 16, 25, \ldots\}$ of perfect squares.

   (e) Compute the the natural density of the set

$$\{1, 3, 4, 5, 7, 9, 11, 12, 13, 15, 16, 17, 19, 20, \ldots\}$$

   of numbers whose binary expansion ends with evenly many zeros.

   Solution.

   (a) We need to break down

$$\frac{\#(A \cap [n])}{n}$$

   Note: $A \cap [n] = \{$ all $n \in [n]$ s.t. $n \in A\}$, so $\#(A \cap [n]) =$ the number of elements from $[n]$ that are also in A.
   Since there are exactly n elements in $[n]$, this implies

$$\frac{\#(A \cap [n])}{n}$$

   is the probability that a randomly selected $b \in [n]$ is in A. This is because randomly selecting b from $[n]$ ensures that all $b \in [n]$ have an equal probability of being selected.

   (b) We want to compute

$$\lim_{n\to\infty} \frac{\#(2\mathbb{Z}_{>0} \cap [n])}{n}$$

   So we need to find some way to express $\#(2\mathbb{Z}_{>0} \cap [n])$ in terms of only n.
   Claim: $\#(2\mathbb{Z}_{>0} \cap [n]) = \lfloor \frac{n}{2} \rfloor$ for all $n \in \mathbb{N}$.
   Proof. We apply mathematical induction on n.
   Base Case: $n = 1$. There are 0 positive even integers in $[n] = [1]$, so

$$\#(2\mathbb{Z}_{>0} \cap [n]) = 0 = \lfloor \frac{1}{2} \rfloor = \lfloor \frac{n}{2} \rfloor$$

   so the claim holds for the base case.
   Inductive Hypothesis: Assume that $\#(2\mathbb{Z}_{>0} \cap [n]) = \lfloor \frac{n}{2} \rfloor$ for all $1 \leq n \leq k$.
   Inductive Step: Consider $n = k + 1$.
   Case 1: n is odd. By the inductive hypothesis, we know there were $\lfloor \frac{k}{2} \rfloor$ positive even integers in $[k]$. Since n is odd, k is even, so $\lfloor \frac{k}{2} \rfloor = \frac{k}{2}$. Also since n is odd, we know

$$\#(2\mathbb{Z}_{>0} \cap [n]) = \#(2\mathbb{Z}_{>0} \cap [k]) = \frac{k}{2}$$

Thus it suffices to show $\frac{k}{2} = \lfloor \frac{n}{2} \rfloor$. Since k is even, we know $\exists\ s \in \mathbb{N}$ s.t. $k = 2s$, which implies $n = 2s + 1$, so we know

$$\lfloor \frac{n}{2} \rfloor = \lfloor \frac{2s+1}{2} \rfloor = s = \frac{k}{2}$$

which completes the proof of Case 1.
Case 2: n is even. By the inductive hypothesis, we know there were $\lfloor \frac{k}{2} \rfloor$ positive even integers in $[k]$. Since n is even, we know that

$$\#(2\mathbb{Z}_{>0} \cap [n]) = \#(2\mathbb{Z}_{>0} \cap [k]) + 1 = \lfloor \frac{k}{2} \rfloor + 1$$

Since n is even, k is odd, so $\lfloor \frac{k}{2} \rfloor = \frac{k-1}{2} = \frac{k+1}{2} - 1 = \frac{n}{2} - 1$. Therefore, we know that

$$\#(2\mathbb{Z}_{>0} \cap [n]) = \#(2\mathbb{Z}_{>0} \cap [k]) + 1 = \lfloor \frac{k}{2} \rfloor + 1 = \frac{n}{2} - 1 + 1 = \frac{n}{2}$$

Since n is even, we know $\frac{n}{2} = \lfloor \frac{n}{2} \rfloor$, which completes the proof that

$$\#(2\mathbb{Z}_{>0} \cap [n]) = \lfloor \frac{n}{2} \rfloor$$

for Case 2.
The conclusion that

$$\#(2\mathbb{Z}_{>0} \cap [n]) = \lfloor \frac{n}{2} \rfloor$$

for all $n \in \mathbb{N}$ follows by induction.
By the definition of the floor function, we know

$$\frac{n-1}{2} \leq \#(2\mathbb{Z}_{>0} \cap [n]) = \lfloor \frac{n}{2} \rfloor \leq \frac{n}{2}$$

This allows us to conclude that

$$\lim_{n\to\infty} \frac{n-1}{2n} \leq \lim_{n\to\infty} \frac{\#(2\mathbb{Z}_{>0} \cap [n])}{n} \leq \lim_{n\to\infty} \frac{n}{2n}$$

Applying L'Hopital's rule to both the upper and lower bound limits, we find

$$\lim_{n\to\infty} \frac{n-1}{2n} = \lim_{n\to\infty} \frac{1}{2} = \frac{1}{2} = \lim_{n\to\infty} \frac{n}{2n}$$

Therefore,

$$\frac{1}{2} \leq \lim_{n\to\infty} \frac{\#(2\mathbb{Z}_{>0} \cap [n])}{n} \leq \frac{1}{2} \implies \lim_{n\to\infty} \frac{\#(2\mathbb{Z}_{>0} \cap [n])}{n} = \frac{1}{2}$$

Thus, we have computed that the natural density of the set $2\mathbb{Z}_{>0}$ is $\frac{1}{2}$.

(c) Let $A =$ the set of positive odd integers. We want to compute

$$\lim_{n\to\infty} \frac{\#(A \cap [n])}{n}$$

We will do so very similarly to part (b). We need to find some way to express $\#(A \cap [n])$ in terms of only n.
Claim: $\#(A \cap [n]) = \lceil \frac{n}{2} \rceil$ for all $n \in \mathbb{N}$.
Proof. We apply mathematical induction on n.
Base Case: $n = 1$. There is one odd positive integer in $[n] = [1]$, so $\#(A \cap [n]) = 1 = \lceil \frac{1}{2} \rceil = \lceil \frac{n}{2} \rceil$, so the claim holds for the base case.
Inductive Hypothesis: Assume $\#(A \cap [n]) = \lceil \frac{n}{2} \rceil$ for all $1 \leq n \leq k$.

Inductive Step: Consider $n = k + 1$.
Case 1: n is odd. By the inductive hypothesis, we know that

$$\#(A \cap [k]) = \lceil \frac{k}{2} \rceil$$

Since n is odd, we know that

$$\#(A \cap [n]) = \#(A \cap [k]) + 1 = \lceil \frac{k}{2} \rceil + 1$$

Also since n is odd, k is even, so $\lceil \frac{k}{2} \rceil = \frac{k}{2}$. Therefore, we know that

$$\#(A \cap [n]) = \frac{k}{2} + 1$$

Since n is odd, $\lceil \frac{n}{2} \rceil = \frac{n+1}{2} = \frac{k+2}{2} = \frac{k}{2} + 1$. Therefore, we have shown that

$$\#(A \cap [n]) = \lceil \frac{n}{2} \rceil$$

which completes Case 1.
Case 2: n is even. By the inductive hypothesis, we know that

$$\#(A \cap [k]) = \lceil \frac{k}{2} \rceil$$

Since n is even, we know that

$$\#(A \cap [n]) = \#(A \cap [k]) = \lceil \frac{k}{2} \rceil$$

Also since n is even, k is odd, so we know $\lceil \frac{k}{2} \rceil = \frac{k+1}{2} = \frac{n}{2}$. Therefore, we know that

$$\#(A \cap [n]) = \frac{n}{2}$$

Since n is even, $\lceil \frac{n}{2} \rceil = \frac{n}{2}$. Therefore, we have shown that

$$\#(A \cap [n]) = \lceil \frac{n}{2} \rceil$$

which completes Case 2.
The conclusion that

$$\#(A \cap [n]) = \lceil \frac{n}{2} \rceil$$

for all $n \in \mathbb{N}$ follows by induction.
By the definition of the ceiling function, we know that

$$\frac{n}{2} \leq \#(A \cap [n]) = \lceil \frac{n}{2} \rceil \leq \frac{n+1}{2}$$

This allows us to conclude that

$$\lim_{n\to\infty} \frac{n}{2n} \leq \lim_{n\to\infty} \frac{\#(A \cap [n])}{n} \leq \lim_{n\to\infty} \frac{n+1}{2n}$$

Applying L'Hopital's rule to both the upper and lower bound limits, we find

$$\lim_{n\to\infty} \frac{n}{2n} = \lim_{n\to\infty} \frac{1}{2} = \frac{1}{2} = \lim_{n\to\infty} \frac{n+1}{2n}$$

Therefore,

$$\frac{1}{2} \leq \lim_{n\to\infty} \frac{\#(A \cap [n])}{n} \leq \frac{1}{2} \implies \lim_{n\to\infty} \frac{\#(A \cap [n])}{n} = \frac{1}{2}$$

Thus, we have computed that the natural density of A = the set of all positive odd integers is $\frac{1}{2}$.

(d) Let $B =$ the set of all perfect squares. Then we want to compute

$$\lim_{n\to\infty} \frac{\#(B \cap [n])}{n}$$

so we need to express $\#(B \cap [n])$ in terms of n.
Claim: $\#(B \cap [n]) = \lfloor \sqrt{n} \rfloor$ for all $n \in \mathbb{N}$.
Proof. We apply mathematical induction on n.
Base Case: $n = 1$. There is one perfect square in the set $[n] = [1]$, so $\#(B \cap [n]) = 1 = \lfloor 1 \rfloor = \lfloor \sqrt{1} \rfloor = \lfloor \sqrt{n} \rfloor$, so the claim holds for the base case.
Inductive Hypothesis: Assume $\#(B \cap [n]) = \lfloor \sqrt{n} \rfloor$ for all $1 \leq n \leq k$.
Inductive Step: Consider $n = k + 1$. By the inductive hypothesis, we know

$$\#(B \cap [k]) = \lfloor \sqrt{k} \rfloor$$

Case 1: n is a perfect square. This implies that

$$\#(B \cap [n]) = \#(B \cap [k]) + 1 = \lfloor \sqrt{k} \rfloor + 1$$

Since $n = i^2$ for some $i \in \mathbb{N}$, we know

$$(i-1)^2 \leq k < i^2 = n \implies i - 1 \leq \sqrt{k} < i \implies \lfloor \sqrt{k} \rfloor = i - 1$$

Therefore, we know that

$$\#(B \cap [n]) = \#(B \cap [k]) + 1 = \lfloor \sqrt{k} \rfloor + 1 = i - 1 + 1 = i$$

Since $n = i^2$, $\sqrt{n} = i$, which completes the proof that

$$\#(B \cap [n]) = \sqrt{n}$$

for Case 1.
Case 2: n is not a perfect square. This implies that

$$\lfloor \sqrt{n-1} \rfloor = \lfloor \sqrt{n} \rfloor$$

Since $n = k + 1$, $n - 1 = k$, so

$$\lfloor \sqrt{n} \rfloor = \lfloor \sqrt{k} \rfloor$$

Combining this with the inductive hypothesis, we find that

$$\#(B \cap [n]) = \lfloor \sqrt{k} \rfloor = \lfloor \sqrt{n} \rfloor$$

which completes the proof for Case 2.
The conclusion that $\#(B \cap [n]) = \lfloor \sqrt{n} \rfloor$ for all $n \in \mathbb{N}$ follows by induction.
By the definition of the floor function, we know

$$\sqrt{n} - 1 \leq \#(B \cap [n]) = \lfloor \sqrt{n} \rfloor \leq \sqrt{n}$$

This allows us to conclude that

$$\lim_{n\to\infty} \frac{\sqrt{n} - 1}{n} \leq \lim_{n\to\infty} \frac{\#(B \cap [n])}{n} \leq \lim_{n\to\infty} \frac{\sqrt{n}}{n}$$

Simplifying known limits, we find that

$$\lim_{n\to\infty} \frac{\sqrt{n} - 1}{n} = \lim_{n\to\infty} \frac{1}{\sqrt{n}} - \lim_{n\to\infty} \frac{1}{n} = \lim_{n\to\infty} \frac{1}{\sqrt{n}} = 0 = \lim_{n\to\infty} \frac{\sqrt{n}}{n}$$

This implies that

$$\lim_{n\to\infty} \frac{\#(B \cap [n])}{n} = \lim_{n\to\infty} \frac{1}{\sqrt{n}} = 0$$

Thus, we have computed that the natural density of $B =$ the set of all perfect squares is 0.

(e) Let C = the set of all positive integers whose binary expansion ends with evenly many 0s.
We want to compute

$$\lim_{n\to\infty} \frac{\#(C \cap [n])}{n}$$

Let C_i = the set of all positive integers whose binary expansion ends with exactly $2i$ 0s for all positive integers i.
Then

$$C = \bigcup_{i=0}^{\infty} C_i \implies \#(C \cap [n]) = \#((\bigcup_{i=0}^{\infty} C_i) \cap [n])$$

Since the C_i's are all mutually disjoint, we know

$$\#(C \cap [n]) = \#((\bigcup_{i=0}^{\infty} C_i) \cap [n] = \sum_{i=0}^{\infty} \#(C_i \cap [n])$$

This implies that

$$\lim_{n\to\infty} \frac{\#(C \cap [n])}{n} = \lim_{n\to\infty} \sum_{i=0}^{\infty} \frac{\#(C_i \cap [n])}{n} = \sum_{i=0}^{\infty} \lim_{n\to\infty} \frac{\#(C_i \cap [n])}{n} \qquad (1)$$

For each $c \in C_i$, the binary expansion of c ends with exactly $2i$ 0s, so the $(2i+1)$'th least significant digit must be 1 in the binary expansion. This digit corresponds to 2^{2i}, and we know that all digits corresponding to 2^k for all $k < 2i$ are 0s. Thus, for all $c \in C_i$,

$$c \equiv 2^{2i} \pmod{2^{2i+1}}$$

The set of possible remainders when dividing by 2^{2i+1} is $\{0, 1, 2, ..., 2^{2i+1} - 1\}$. There are 2^{2i+1} total possible remainders, only one of which is 2^{2i}. Therefore, we know that a $c \in C_i$ appears exactly once for every 2^{2i+1} positive integers. This implies that

$$\#(C_i \cap [n]) = \lceil \frac{n}{2^{2i+1}} \rceil$$

By the definition of the ceiling function, we know

$$\frac{n}{2^{2i+1}} \leq \#(C_i \cap [n]) = \lceil \frac{n}{2^{2i+1}} \rceil \leq \frac{n}{2^{2i+1}} + 1$$

This allows us to conclude that

$$\lim_{n\to\infty} \frac{n}{n2^{2i+1}} = \frac{1}{2^{2i+1}} \leq \lim_{n\to\infty} \frac{\#(C_i \cap [n])}{n} \leq \frac{1}{2^{2i+1}} = \lim_{n\to\infty} \frac{n}{n(2^{2i+1})} + \frac{1}{n}$$

So we know that

$$\lim_{n\to\infty} \frac{\#(C_i \cap [n])}{n} = \frac{1}{2^{2i+1}}$$

Plugging this into (1), we find that

$$\lim_{n\to\infty} \frac{\#(C \cap [n])}{n} = \sum_{i=0}^{\infty} \lim_{n\to\infty} \frac{\#(C_i \cap [n])}{n} = \sum_{i=0}^{\infty} \frac{1}{2^{2i+1}} = \frac{1}{2} \sum_{i=0}^{\infty} \frac{1}{4^i} = \frac{1}{2} \frac{1}{1 - \frac{1}{4}} = \frac{1}{2} \frac{1}{\frac{3}{4}} = \frac{1}{2} \frac{4}{3} = \frac{2}{3}$$

Thus, we have computed that the density of C = the set of all positive integers whose binary expansion ends with an even number of 0s is $\frac{2}{3}$.

5. (Ross P3.18) In a certain community, 36 percent of the families own a dog and 22 percent of the families that own a dog also own a cat. In addition, 30 percent of the families own a cat. What is

   (a) the probability that a randomly selected family owns both a dog and a cat?

   (b) the conditional probability that a randomly selected family owns a dog given that it owns a cat?

Solution.
First, we will define notation to represent the given information.
Let D = the event that a randomly selected family owns a dog.
Let C = the event that a randomly selected family owns a cat.
Then we are given the following information: $\begin{cases} \mathbb{P}(D) = 0.36 \\ \mathbb{P}(C|D) = 0.22 \\ \mathbb{P}(C) = 0.30 \end{cases}$

(a) We want to find $\mathbb{P}(CD)$. By the definition of conditional probability, we know

$$\mathbb{P}(C|D) = \frac{\mathbb{P}(CD)}{\mathbb{P}(D)} \qquad (1)$$

Plugging known values into (1), we find

$$0.22 = \frac{\mathbb{P}(CD)}{0.36} \implies \mathbb{P}(CD) = 0.22 \cdot 0.36 = 0.0792 = 7.92\%$$

Therefore, the probability that a randomly selected family owns both a dog and a cat is 7.92%.

(b) We want to find $\mathbb{P}(D|C)$. By Bayes' Theorem, we know that

$$\mathbb{P}(D|C) = \frac{\mathbb{P}(C|D) \cdot \mathbb{P}(D)}{\mathbb{P}(C)} = \frac{0.22 \cdot 0.36}{0.30} = \frac{0.0792}{0.30} = 0.264 = 26.40\%$$

Thus, the probability that a randomly selected family owns a dog given that it owns a cat is 26.40%.

6. (Ross P3.24) Urn I contains 2 white and 4 red balls, whereas urn II contains 1 white and 1 red ball. A ball is randomly chosen from urn I and put into urn II, and a ball is then randomly selected from urn II. What is

   (a) the probability that the ball selected from urn II is white?

   (b) the conditional probability that the transferred ball was white given that a white ball is selected from urn II?

   Solution.

   (a) The ball that is randomly selected from urn I and put into urn II can either be white or red.
   Let b_1 = the ball that is randomly selected from urn I and put into urn II.
   Let b_2 = the ball that is then randomly selected from urn II.
   Let R_1 = the event that b_1 is red.
   Let W_1 = the event that b_1 is white.
   Let W_2 = the even that b_2 is white.
   Since b_1 is randomly selected from urn I, there is an equal probability of each ball in urn I being selected. There are $4 + 2 = 6$ total balls in urn I. Two of these are white, and there is an equal probability of selecting each ball, so

$$\mathbb{P}(W_1) = \frac{2}{6} = \frac{1}{3} \approx 33.33\%$$

   The remaining four balls from urn I are red, and there is an equal probability of selecting each ball, so

$$\mathbb{P}(R_1) = \frac{4}{6} = \frac{2}{3} \approx 66.67\%$$

   Since b_1 must be either red or white, the Law of Total Probability guarantees that

$$\mathbb{P}(W_2) = \mathbb{P}(W_2|R_1)\mathbb{P}(R_1) + \mathbb{P}(W_2|W_1)\mathbb{P}(W_1) \qquad (1)$$

   We already calculated $\mathbb{P}(W_1)$ and $\mathbb{P}(R_1)$, so we just need to compute $\mathbb{P}(W_2|R_1)$ and $\mathbb{P}(W_2|W_1)$.
   For $\mathbb{P}(W_2|R_1)$, we are given that b_1 is red, so we know that urn II has 2 red balls and 1 white ball. Thus, the probability that b_2 is white given that b_1 is red is

$$\mathbb{P}(W_2|R_1) = \frac{1}{3} \approx 33.33\%$$

   For $\mathbb{P}(W_2|W_1)$, we are given that b_1 is white, so we know that urn II has 1 red ball and 2 white balls. Thus, the probability that b_2 is white given that b_1 is white is

$$\mathbb{P}(W_2|W_1) = \frac{2}{3} \approx 66.67\%$$

   Plugging $P(W_2|R_1)$, $\mathbb{P}(W_2|W_1)$, $\mathbb{P}(W_1)$, and $\mathbb{P}(R_1)$ into (1), we find

$$\mathbb{P}(W_2) = \frac{1}{3}\frac{2}{3} + \frac{2}{3}\frac{1}{3} = 2\frac{2}{9} = \frac{4}{9} \approx 44.44\%$$

   (b) We will use the same definitions as in part (a). Then we want to find

$$\mathbb{P}(W_1|W_2)$$

   Bayes' Theorem guarantees that

$$\mathbb{P}(W_1|W_2) = \frac{\mathbb{P}(W_2|W_1)\mathbb{P}(W_1)}{\mathbb{P}(W_2)}$$

In part (a), we calculated that $\begin{cases} \mathbb{P}(W_2|W_1) = \frac{2}{3} \\ \mathbb{P}(W_1) = \frac{1}{3} \\ \mathbb{P}(W_2) = \frac{4}{9} \end{cases}$ Therefore, we can easily compute that the conditional probability that the transferred ball was white given that a white ball is selected from urn II is

$$\mathbb{P}(W_1|W_2) = \frac{\frac{2}{3}\frac{1}{3}}{\frac{4}{9}} = \frac{2}{9}\frac{9}{4} = \frac{2}{4} = \frac{1}{2} = 50\%$$

7. The *Riemann zeta function* is defined to be

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

This series converges for $s > 1$ (indeed, for s complex with real part > 1), though we will ignore such technicalities here.

Leonhard Euler famously solved the *Basel problem* in 1734 by showing that

$$\zeta(2) = 1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots = \frac{\pi^2}{6},$$

a remarkable calculation which shows up in many unexpected places.

In this problem, your task is to instead justify Euler's product formula for the zeta function. Specifically, show that

$$\zeta(s) = \prod_p \frac{1}{1 - p^{-s}},$$

where the product is over all prime numbers $p = 2, 3, 5, \ldots$.

Hint: Expand the fraction as a geometric series. You do not need to provide rigorous justification for your manipulations (Euler didn't!) so long as they are "plausible." An appropriate course in real or complex analysis will supply such missing details.

Solution.
We want to show that

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_p \frac{1}{1 - p^{-s}}$$

Note:

$$\prod_p \frac{1}{1 - p^{-s}} = \prod_p \frac{1}{1 - \frac{1}{p^s}}$$

Since p is prime and $s > 1$, we know $\frac{1}{p^s} < 1$, so the geometric series identity guarantees that

$$\prod_p \frac{1}{1 - \frac{1}{p^s}} = \prod_p \sum_{i=0}^{\infty} \frac{1}{(p^s)^i}$$

Expanding the infinite product, we find

$$\begin{aligned} \prod_p \frac{1}{1 - \frac{1}{p^s}} &= \sum_{i=0}^{\infty} \frac{1}{(2^s)^i} \cdot \sum_{i=0}^{\infty} \frac{1}{(3^s)^i} \cdot \sum_{i=0}^{\infty} \frac{1}{(5s)^i} \cdot \cdots \\ &= (1 + \frac{1}{2^s} + \frac{1}{(2^s)^2} + \cdots) \cdot (1 + \frac{1}{3^s} + \frac{1}{(3^s)^2} + \cdots) \cdot (1 + \frac{1}{5^s} + \frac{1}{(5^s)^2} + \cdots) \cdot \cdots \quad (1) \end{aligned}$$

Observation 1: For each prime p and for each $0 \leq i \in \mathbb{Z}$, $\frac{1}{(p^s)^i}$ appears exactly once in some term of the above product. Therefore, in each term t of the infinite sum that results from expanding the product, if $\frac{1}{(p^s)^i}$ is a factor of t, then $\sum_{i=0}^{\infty} \frac{1}{(p^s)^i}$ must have contributed that factor.
Observation 2: For each prime p and each term t (defined above), the sum $\sum_{i=0}^{\infty} \frac{1}{(p^s)^i}$ contributes exactly one factor to t. Therefore, if a term t has $\frac{1}{(p^s)^i}$ as a factor, then the term which $\sum_{i=0}^{\infty} \frac{1}{(p^s)^i}$ contributed must be $\frac{1}{(p^s)^i}$ itself, not some combination of $\frac{1}{(p^s)^j}$ and $\frac{1}{(p^s)^k}$, where $k + j = i$.

Combining these two observations, we can see that expanding the infinite product will produce exactly one term $t = \frac{1}{(p_1^s)^{\alpha_1}(p_2^s)^{\alpha_2}(p_3^s)^{\alpha_3}\cdots}$ for all primes p and for all $0 \leq \alpha_1, \alpha_2, \alpha_3, \cdots \in \mathbb{Z}$.
Let

$$t_1 = \frac{1}{(p_1^s)^{\alpha_1}(p_2^s)^{\alpha_2}(p_3^s)^{\alpha_3}\cdots}$$

and

$$t_2 = \frac{1}{(p_1^s)^{\beta_1}(p_2^s)^{\beta_2}(p_3^s)^{\beta_3}\cdots}$$

Then, by the Fundamental Theorem of Arithmetic, we know that

$$t_1 = t_2 \iff \alpha_i = \beta_i \quad \forall i$$

For all $\alpha_i \geq 1$, $t = \frac{1}{(p_1^s)^{\alpha_1}(p_2^s)^{\alpha_2}(p_3^s)^{\alpha_3}\cdots} = \frac{1}{(p_1^{\alpha_1})^s(p_2^{\alpha_2})^s(p_3^{\alpha_3})^s\cdots} = \left(\frac{1}{p_1^{\alpha_1}p_2^{\alpha_2}p_3^{\alpha_3}\cdots}\right)^s$ appears exactly once in the expansion of (1), so we know that all terms t in the expansion of (1) are distinct. Since the Fundamental Theorem of Arithmetic guarantees that all natural numbers can be written uniquely as the product of primes, this implies that $t = \frac{1}{n^s}$ appears exactly once in the expansion of (1) for all $n \in \mathbb{N}$. Rearranging terms in the expansion of (1), we see that

$$\begin{aligned}
\prod_p \frac{1}{1-p^{-s}} = &\quad \sum_{i=0}^{\infty}\frac{1}{(2^s)^i}\cdot\sum_{i=0}^{\infty}\frac{1}{(3^s)^i}\cdot\sum_{i=0}^{\infty}\frac{1}{(5s)^i}\cdots \\
= &\quad (1+\frac{1}{2^s}+\frac{1}{(2^s)^2}+\cdots)\cdot(1+\frac{1}{3^s}+\frac{1}{(3^s)^2}+\cdots)\cdot(1+\frac{1}{5^s}+\frac{1}{(5^s)^2}+\cdots)\cdots \\
= &\quad \frac{1}{(2^0 3^0 5^0\cdots)^s}+\frac{1}{(2^1 3^0 5^0\cdots)^s}+\frac{1}{(2^0 3^1 5^0\cdots)^s}+\frac{1}{(2^2 3^0 5^0\cdots)^s}+ \\
&\quad \frac{1}{(2^0 3^0 5^1 7^0\cdots)^s}+\cdots \\
= &\quad \frac{1}{1^s}+\frac{1}{2^s}+\frac{1}{3^s}+\frac{1}{4^s}+\frac{1}{5^s}+\cdots \\
= &\quad \sum_{n=1}^{\infty}\frac{1}{n^s} = \zeta(s)
\end{aligned}$$

This completes the proof that

$$\zeta(s) = \sum_{n=1}^{\infty}\frac{1}{n^s} = \prod_p \frac{1}{1-p^{-s}}$$

Note: p_i refers to the ith prime number, where $p_1 = 2$.

8. A number is *squarefree* if it is a product of distinct primes. The set of squarefree numbers is

$$S = \{1, 2, 3, 5, 6, 7, 10, 11, 13, \ldots\}.$$

In this problem, you will follow an outline for a proof of the classic result that the natural density of the squarefree numbers is $6/\pi^2 \approx 0.607927$. Sometimes this result is informally stated as "most numbers are squarefree."

(a) Write p_i for the ith prime number, so $p_1, p_2, p_3, \ldots = 2, 3, 5, \ldots$. Let

$$E_i = \{p_i^2, 2p_i^2, 3p_i^2, \ldots\}$$

be the set of multiples of p_i^2. Show that the natural density of E_i is $1/p_i^2$.

Hint: Show that $\#E_i \cap [n] = \left\lfloor \frac{n}{p_i^2} \right\rfloor$ where $\lfloor x \rfloor$ is the *floor* of x, defined to be the unique integer satisfying $\lfloor x \rfloor \leq x < \lfloor x \rfloor + 1$.

(b) Let $T = S^c = \{4, 8, 9, 12, \ldots\}$ be the set of non-squarefree positive integers. Show that

$$T = E_1 \cup E_2 \cup \cdots.$$

More precisely, show that

$$T \cap [n] = \bigcup_{i=1}^{m} E_i \cap [n],$$

where m is the largest index for which $p_i^2 \leq n$.

(c) Show that

$$\#T \cap [n] = \sum_{\varnothing \neq I \subseteq [m]} (-1)^{\#I - 1} \cdot \left\lfloor \frac{n}{\prod_{i \in I} p_i^2} \right\rfloor.$$

Conclude that

$$\#S \cap [n] = \sum_{I \subseteq [m]} (-1)^{\#I} \cdot \left\lfloor \frac{n}{\prod_{i \in I} p_i^2} \right\rfloor.$$

(d) Show that

$$\sum_{I \subseteq [m]} (-1)^{\#I} \frac{1}{\prod_{i \in I} p_i^2} = \prod_{i=1}^{m} \left(1 - \frac{1}{p_i^2}\right).$$

(e) Provide some justification for each of the following steps except (23):

$$\lim_{n \to \infty} \frac{\#S \cap [n]}{n} = \lim_{n \to \infty} \frac{1}{n} \sum_{I \subseteq [m]} (-1)^{\#I} \left\lfloor \frac{n}{\prod_{i \in I} p_i^2} \right\rfloor \tag{22}$$

$$= \lim_{n \to \infty} \frac{1}{n} \sum_{I \subseteq [m]} (-1)^{\#I} \frac{n}{\prod_{i \in I} p_i^2} \tag{23}$$

$$= \prod_{i=1}^{\infty} \left(1 - \frac{1}{p_i^2}\right) \tag{24}$$

$$= \frac{1}{\zeta(2)} \tag{25}$$

$$= \frac{6}{\pi^2}. \tag{26}$$

(f) Explain the result of the calculation in (e) in your own words.

Solution.

(a) We want to show that

$$\lim_{n\to\infty} \frac{\#(E_i \cap [n])}{n} = \frac{1}{p_i^2}$$

Claim: $\#(E_i \cap [n]) = \lfloor \frac{n}{p_i^2} \rfloor$ for all i.

Proof. We apply mathematical induction on n.

Base Case: $n = 1$. For all primes p, $p > 1 \implies p^2 > 1 \implies kp^2 \notin [n] = [1]$ for all $k \in \mathbb{N}$. Thus,

$$\#(E_i \cap [n]) = \#(E_i \cap [1]) = 0$$

Also for all primes p, $p^2 > 1 \implies 0 \leq \frac{1}{p^2} < 1$. Thus, $\lfloor \frac{1}{p^2} \rfloor = 0$, so

$$\#(E_i \cap [n]) = \#(E_i \cap [1]) = 0 = \lfloor \frac{1}{p^2} \rfloor = \lfloor \frac{n}{p^2} \rfloor$$

so the claim holds for the base case.

Inductive Hypothesis: Assume $\#(E_i \cap [n]) = \lfloor \frac{n}{p_i^2} \rfloor$ for all i and for all $1 \leq n \leq j$.

Inductive Step: Consider $n = j + 1$. By the inductive hypothesis, we know

$$\#(E_i \cap [j]) = \lfloor \frac{j}{p_i^2} \rfloor$$

for all i.

For any i, one of the following is true:

Case 1: $n = kp_i^2$ for some $k \in \mathbb{N}$. In this case, we know

$$\#(E_i \cap [n]) = \#(E_i \cap [j]) + 1 = \lfloor \frac{j}{p_i^2} \rfloor + 1 \qquad (1)$$

Since $n = kp_i^2$, we know $\frac{n}{p_i^2} = k \in \mathbb{N}$, so

$$\lfloor \frac{n}{p_i^2} \rfloor = \frac{n}{p_i^2}$$

Also, by the definition of the floor function, since

$$\frac{n - p_i^2}{p_i^2} \leq \frac{n-1}{p_i^2} < \frac{n - p_i^2}{p_i^2} + 1 = \lfloor \frac{n}{p_i^2} \rfloor$$

we know that

$$\lfloor \frac{n-1}{p_i^2} \rfloor = \lfloor \frac{j}{p_i^2} \rfloor = \frac{n - p_i^2}{p_i^2}$$

Plugging this into (1), we find

$$\#(E_i \cap [n]) = \#(E_i \cap [j]) + 1 = \lfloor \frac{j}{p_i^2} \rfloor + 1 = \frac{n - p_i^2}{p_i^2} + 1 = \frac{n}{p_i^2} = \lfloor \frac{n}{p_i^2} \rfloor$$

which completes the proof for Case 1.

Case 2: $n \neq kp_i^2$ for all $k \in \mathbb{N}$. Then we know

$$\#(E_i \cap [n]) = \#(E_i \cap [j]) = \lfloor \frac{j}{p_i^2} \rfloor$$

Suppose $j \equiv x \pmod{p_i^2}$ and $n \equiv x + 1 \pmod{p_i^2}$. Then $0 \leq x < p_i^2 - 1$ and $1 \leq x + 1 \leq p_i^2 - 1$.

Clearly, $\frac{j-x}{p_i^2} \in \mathbb{Z}$, and

$$\frac{j-x}{p_i^2} \leq \frac{j}{p_i^2} = \frac{j-x}{p_i^2} + \frac{x}{p_i^2} < \frac{j-x}{p_i^2} + \frac{p_i^2}{p_i^2} = \frac{j-x}{p_i^2} + 1$$

By the definition of the floor function, this implies

$$\lfloor \frac{j}{p_i^2} \rfloor = \frac{j-x}{p_i^2}$$

And we also know

$$\frac{j-x}{p_i^2} = \frac{n-(x+1)}{p_i^2} \leq \frac{n}{p_i^2} = \frac{n-(x+1)}{p_i^2} + \frac{x+1}{p_i^2} < \frac{n-(x+1)}{p_i^2} + \frac{p_i^2}{p_i^2} = \frac{j-x}{p_i^2} + 1$$

So by the definition of the floor function, we know

$$\lfloor \frac{n}{p_i^2} \rfloor = \frac{n-(x+1)}{p_i^2} = \frac{j-x}{p_i^2} = \lfloor \frac{j}{p_i^2} \rfloor$$

This completes the proof that

$$\#(E_i \cap [n]) = \#(E_i \cap [j]) = \lfloor \frac{j}{p_i^2} \rfloor = \lfloor \frac{n}{p_i^2} \rfloor$$

for Case 2.
The conclusion that

$$\#(E_i \cap [n]) = \lfloor \frac{n}{p_i^2} \rfloor$$

for all $n \in N$ and all i follows by induction.
This implies that

$$\frac{n}{p_i^2} - 1 \leq \#(E_i \cap [n]) = \lfloor \frac{n}{p_i^2} \rfloor \leq \frac{n}{p_i^2}$$

so we know that

$$\lim_{n\to\infty} \frac{\frac{n}{p_i^2} - 1}{n} \leq \lim_{n\to\infty} \frac{\#(E_i \cap [n])}{n} \leq \lim_{n\to\infty} \frac{n}{np_i^2}$$

Simplifying both the upper and lower bound limits, we find that

$$\lim_{n\to\infty} \frac{\frac{n}{p_i^2} - 1}{n} = \lim_{n\to\infty} \frac{1}{p_i^2} - \lim_{n\to\infty} \frac{1}{n} = \frac{1}{p_i^2} = \lim_{n\to\infty} \frac{n}{np_i^2}$$

This directly implies that

$$\lim_{n\to\infty} \frac{\#(E_i \cap [n])}{n} = \frac{1}{p_i^2}$$

for all i, which completes the proof that the natural density of E_i is $\frac{1}{p_i^2}$.

(b) We want to show that

$$T \cap [n] = \bigcup_{i=1}^{m} E_i \cap [n] \qquad (2)$$

where m is the largest index for which $p_i^2 \leq n$.
It suffices to show that for all $t \in T \cap [n]$, $t \in \bigcup_{i=1}^m E_i \cap [n]$, and for all $e \in \bigcup_{i=1}^m E_i \cap [n]$, $e \in T \cap [n]$.
Note: for all i, $E_i \cap [n] = \{e \in E_i | e \leq n\}$.
Similarly, $T \cap [n] = \{t \in T | t \leq n\}$.

First, we will show that for all $t \in T \cap [n]$, $t \in \bigcup_{i=1}^m E_i \cap [n]$.
Let $t = p_y^2 s$ where $s \in \mathbb{N}$ and p_y is the largest prime whose square divides t.

$$p_y^2 | t \implies p_y^2 \leq t \implies 1 \leq y \leq m$$

so we know $E_y \cap [n]$ is a term in the union from (2). Also, since

$$E_y \cap [n] = \{e \in E_y | e \leq n\}$$

and $p_y^2 s = t \leq n$, we know $p_y^2 s \in E_y \cap [n]$ for all $t \in T \cap [n]$.

Now, we will show that for all $e \in \bigcup_{i=1}^m E_i \cap [n]$, $e \in T \cap [n]$. Consider an arbitrary $e \in \bigcup_{i=1}^m E_i \cap [n]$. Then $\exists\ 1 \leq z \leq m$ s.t. $e \in E_z \cap [n]$. Therefore, we know that $1 \leq e \leq n$ and $e = ap_z^2$, so e must be a non-squarefree positive integer. Thus

$$e \in T$$

and we know $e \leq n$, so we know

$$e \in T \cap [n]$$

Thus, we have shown that, for all $t \in T \cap [n]$, $t \in \bigcup_{i=1}^m E_i \cap [n]$, and for all $e \in \bigcup_{i=1}^m E_i \cap [n]$, $e \in T \cap [n]$.
This completes the proof that

$$T \cap [n] = \bigcup_{i=1}^{m} E_i \cap [n]$$

where m is the largest index for which $p_i^2 \leq n$.

(c) From part (b), we know that

$$T \cap [n] = \bigcup_{i=1}^{m} E_i \cap [n]$$

By the Principle of Inclusion Exclusion, we know that

$$\#(T \cap [n]) = \sum_{1 \leq i_1 < \ldots < i_k \leq m} \#\Big((E_{i_1} \cap [n]) \cap \cdots \cap (E_{i_k} \cap [n])\Big) = \sum_{\emptyset \neq I \subseteq [m]} (-1)^{\#I-1} \#(\bigcap_{i \in I} (E_i \cap [n])) \qquad (3)$$

From part (a), we know that

$$\#(E_i \cap [n]) = \lfloor \frac{n}{p_i^2} \rfloor$$

for all i. In every p_i^2 consecutive positive integers, exactly one of them will be divisible by p_i^2, so over the n consecutive positive integers in $[n]$, there are $\lfloor \frac{n}{p_i^2} \rfloor$ integers which are multiples of p_i^2 (and thus elements of E_i).
Similarly, for any $e \in \bigcap_{i \in I} (E_i \cap [n])$, e must be a multiple of p_i^2 for all $i \in I$. Since all primes p, j have $gcd(p, j) = 1$, such an e only appears once in every $\prod_{i \in I} p_i^2$ consecutive positive integers. Thus, over the n consecutive positive integers in $[n]$, there are maximally $\frac{n}{\prod_{i \in I} p_i^2}$ integers which belong to $\bigcap_{i \in I} (E_i \cap [n])$. If $n \notin \bigcap_{i \in I} (E_i \cap [n])$, then $\frac{n}{\prod_{i \in I} p_i^2} \notin \mathbb{N}$, so we can only have $\lfloor \frac{n}{\prod_{i \in I} p_i^2} \rfloor$ elements in $\bigcap_{i \in I} (E_i \cap [n])$. Therefore, in either case, the number of elements in $\bigcap_{i \in I} (E_i \cap [n])$ is exactly

$$\lfloor \frac{n}{\prod_{i \in I} p_i^2} \rfloor$$

Plugging this into (3), we find that

$$\#(T \cap [n]) = \sum_{\emptyset \neq I \subseteq [m]} (-1)^{\#I-1} \#(\bigcap_{i \in I} (E_i \cap [n])) = \sum_{\emptyset \neq I \subseteq [m]} (-1)^{\#I-1} \lfloor \frac{n}{\prod_{i \in I} p_i^2} \rfloor \qquad (4)$$

as required. To arrive at the final conclusion for part (c), note that all integers from 1 to n are either squarefree or are not, so

$$|T \cap [n]| + |S \cap [n]| = n \implies |S \cap [n]| = n - |T \cap [n]$$

Plugging (4) into this equation, we find that

$$\#(S \cap [n]) = n - \sum_{\emptyset \neq I \subseteq [m]} (-1)^{\#I-1} \lfloor \frac{n}{\prod_{i \in I} p_i^2} \rfloor = n + \sum_{\emptyset \neq I \subseteq [m]} (-1)^{\#I} \lfloor \frac{n}{\prod_{i \in I} p_i^2} \rfloor$$

Note: If $I = \emptyset$, then

$$(-1)^{\#I} \# \lfloor \frac{n}{\prod_{i \in I} p_i^2} \rfloor = (-1)^0 \frac{n}{1} = \frac{n}{1} = n$$

Thus,

$$n + \sum_{\emptyset \neq I \subseteq [m]} (-1)^{\#I} \#(\bigcap_{i \in I} \lfloor \frac{n}{\prod_{i \in I} p_i^2} \rfloor = \sum_{I \subseteq [m]} (-1)^{\#I} \lfloor \frac{n}{\prod_{i \in I} p_i^2} \rfloor$$

so we know

$$\#(S \cap [n]) = \sum_{I \subseteq [m]} (-1)^{\#I} \lfloor \frac{n}{\prod_{i \in I} p_i^2} \rfloor$$

which completes part (c).

(d) We want to show

$$\sum_{I \subseteq [m]} (-1)^{\#I} \frac{1}{\prod_{i \in I} p_i^2} = \prod_{i=1}^{m} \left(1 - \frac{1}{p_i^2}\right) \tag{5}$$

We know that

$$\sum_{I \subseteq [m]} (-1)^{\#I} \frac{1}{\prod_{i \in I} p_i^2} = 1 - \sum_{\emptyset \neq I \subseteq [m]} (-1)^{\#I-1} \frac{1}{\prod_{i \in I} p_i^2}$$

The probability that a randomly selected $b \in [n]$ is not squarefree is

$$\mathbb{P}(b \in T | b \in [n]) = \mathbb{P}(\bigcup_{i=1}^{m} E_i)$$

Since the E_i's are not mutually disjoint, we can apply the Principle of Inclusion Exclusion to find that

$$\mathbb{P}(b \in T | b \in [n]) = \mathbb{P}(\bigcup_{i=1}^{m} E_i) = \sum_{k=1}^{m} (-1)^{k-1} \sum_{1 \leq i_1 < \cdots < i_k \leq m} \mathbb{P}(E_{i_1} \cap \cdots \cap E_{i_k}) \tag{6}$$

For any $e \in E_{i_1} \cap \cdots \cap E_{i_k}$, $p_{i_1}^2 \cdot \cdots \cdot p_{i_k}^2 | e$. This only happens once for every $p_{i_1}^2 \cdot \cdots \cdot p_{i_k}^2$ consecutive positive integers, so the probability that a randomly selected $b \in [n]$ is divisible by $p_{i_1}^2 \cdot \cdots \cdot p_{i_k}^2$ is

$$\mathbb{P}(E_{i_1} \cap \cdots \cap E_{i_k}) = \frac{1}{p_{i_1}^2 \cdot \cdots \cdot p_{i_k}^2} = \frac{1}{\prod_{j=1}^{k} p_{i_j}^2}$$

Plugging this into (6), we find

$$\mathbb{P}(b \in T | b \in [n]) = \sum_{k=1}^{m} (-1)^{k-1} \sum_{1 \leq i_1 < \cdots < i_k \leq m} \frac{1}{\prod_{j=1}^{k} p_{i_j}^2} \tag{7}$$

Since we are summing over sets $\{i_1, \cdots, i_k\} \subseteq [m]$ of size k for all k from $1 \to n$, we are summing over all nonempty subsets of $\emptyset \neq I \subseteq [m]$. Therefore, we can rewrite (7) as a singular sum over $\emptyset \neq I \subseteq [m]$, and we find

$$\mathbb{P}(b \in T | b \in [n]) = \sum_{\emptyset \neq I \subseteq [m]} (-1)^{\#I-1} \frac{1}{\prod_{i \in I} p_i^2}$$

Since $S = T^c$, $(S \cap [n]) = (T \cap [n])^c$, so $\mathbb{P}(b \in S | b \in [n]) = 1 - \mathbb{P}(b \in T | b \in [n])$, so we know that

$$\mathbb{P}(b \in S | b \in [n]) = 1 - \sum_{\emptyset \neq I \subseteq [m]} (-1)^{\#I - 1} \frac{1}{\prod_{i \in I} p_i^2}$$

We already know that

$$\sum_{I \subseteq [m]} (-1)^{\#I} \frac{1}{\prod_{i \in I} p_i^2} = 1 - \sum_{\emptyset \neq I \subseteq [m]} (-1)^{\#I - 1} \frac{1}{\prod_{i \in I} p_i^2}$$

so we can conclude that

$$\mathbb{P}(b \in S | b \in [n]) = \sum_{I \subseteq [m]} (-1)^{\#I} \frac{1}{\prod_{i \in I} p_i^2}$$

Now, let's compute this probability in a different way. For a randomly selected integer $b \in [n]$ to be squarefree, it must not have any p_i^2 as a factor for all $1 \leq i \leq m$. From part (a), we know that exactly one out of every p_i^2 consecutive positive integers has p_i^2 as a factor. Therefore, the probability that b has p_i^2 as a factor is

$$\mathbb{P}(p_i^2 | b) = \frac{1}{p_i^2}$$

Taking the complement, we see that the probability that b *does not* have p_i^2 as a factor is

$$\mathbb{P}(p_i^2 \nmid b) = 1 - \frac{1}{p_i^2}$$

This is mutually independently true for all $1 \leq i \leq m$, and so the probability that b has no p_i^2 factors for all $1 \leq i \leq m$, which is the probability that b is squarefree, is

$$\mathbb{P}(b \in S | b \in [n]) = \prod_{i=1}^{m} (1 - \frac{1}{p_i^2})$$

Thus, we have shown that the probability that a randomly selected integer $b \in [n]$ is squarefree is

$$\mathbb{P}(b \in S | b \in [n]) = \sum_{I \subseteq [m]} (-1)^{\#I} \frac{1}{\prod_{i \in I} p_i^2} = \prod_{i=1}^{m} \left(1 - \frac{1}{p_i^2}\right)$$

which completes the proof for part (d).

(e) (1) From part (c), we know that

$$\#(S \cap [n]) = \sum_{I \subseteq [m]} (-1)^{\#I} \lfloor \frac{n}{\prod_{i \in I} p_i^2} \rfloor$$

Multiplying both sides by $\frac{1}{n}$, we obtain

$$\frac{1}{n} \#(S \cap [n]) = \frac{1}{n} \sum_{I \subseteq [m]} (-1)^{\#I} \lfloor \frac{n}{\prod_{i \in I} p_i^2} \rfloor$$

Taking the limit as $n \to \infty$ of both sides, we obtain

$$\lim_{n \to \infty} \frac{1}{n} \#(S \cap [n]) = \lim_{n \to \infty} \frac{1}{n} \sum_{I \subseteq [m]} (-1)^{\#I} \lfloor \frac{n}{\prod_{i \in I} p_i^2} \rfloor$$

which completes the justification of step (1).

(2) No justification requested for step (2).

(3) Simplifying the equation from step (2), we find that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{I\subseteq[m]}(-1)^{\#I}\frac{n}{\prod_{i\in I}p_i^2}=\lim_{n\to\infty}\sum_{I\subseteq[m]}(-1)^{\#I}\frac{1}{\prod_{i\in I}p_i^2}$$

From part (d), we know that

$$\sum_{I\subseteq[m]}(-1)^{\#I}\frac{1}{\prod_{i\in I}p_i^2}=\prod_{i=1}^{m}\left(1-\frac{1}{p_i^2}\right)$$

Taking the limit as $n\to\infty$ on both sides, we obtain

$$\lim_{n\to\infty}\sum_{I\subseteq[m]}(-1)^{\#I}\frac{1}{\prod_{i\in I}p_i^2}=\lim_{n\to\infty}\prod_{i=1}^{m}\left(1-\frac{1}{p_i^2}\right)$$

Note: As $n\to\infty$, $m\to\infty$, as the largest prime s.t. $p_m^2\le n$ increases as n increases. Therefore,

$$\lim_{n\to\infty}\prod_{i=1}^{m}\left(1-\frac{1}{p_i^2}\right)=\prod_{i=1}^{\infty}\left(1-\frac{1}{p_i^2}\right)$$

Therefore, we know that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{I\subseteq[m]}(-1)^{\#I}\frac{n}{\prod_{i\in I}p_i^2}=\prod_{i=1}^{\infty}\left(1-\frac{1}{p_i^2}\right)$$

which completes the justification of step (3).

(4) Note that

$$\prod_{i=1}^{\infty}\left(1-\frac{1}{p_i^2}\right)=\prod_{i=1}^{\infty}\left(1-p_i^{-2}\right)$$

Taking the reciprocal, we find

$$\frac{1}{\prod_{i=1}^{\infty}\left(1-\frac{1}{p_i^2}\right)}=\frac{1}{\prod_{i=1}^{\infty}\left(1-p_i^{-2}\right)}=\prod_{i=1}^{\infty}\frac{1}{1-p_i^{-2}}$$

From question 7, we know that

$$\prod_{i=1}^{\infty}\frac{1}{1-p_i^{-2}}=\sum_{n=1}^{\infty}\frac{1}{n^2}=\zeta(2)$$

Therefore,

$$\prod_{i=1}^{\infty}\left(1-\frac{1}{p_i^2}\right)=\frac{1}{\prod_{i=1}^{\infty}\frac{1}{1-p_i^{-2}}}=\frac{1}{\zeta(2)}$$

which completes the justification for step (4).

(5) From question 7, we know that

$$\zeta(2)=\frac{\pi^2}{6}$$

This directly implies that

$$\frac{1}{\zeta(2)}=\frac{1}{\frac{\pi^2}{6}}=\frac{6}{\pi^2}$$

which completes the justification for step (5).

(f) Since $\#(S \cap [n])$ = the # of squarefree integers in $[n]$, and there are n integers in $[n]$, we can interpret

$$\frac{\#(S \cap [n])}{n}$$

as the probability that a randomly selected integer $b \in [n]$ is squarefree.
As $n \to \infty$, $[n]$ approaches the set of natural numbers, $\mathbb{N}$, and $S \cap [n]$ approaches the set of all squarefree natural numbers. Therefore, we can interpret

$$\lim_{n\to\infty} \frac{\#(S \cap [n])}{n}$$

as the probability that a randomly selected natural number is squarefree. The result of the calculation from part (e) demonstrates that

$$\lim_{n\to\infty} \frac{\#(S \cap [n])}{n} = \frac{6}{\pi^2}$$

Therefore, we can explain the result of the calculation in (e) by saying that the total probability that a randomly selected natural number is squarefree is $\frac{6}{\pi^2} \approx 60.79\%$. One direct implication of this result is that the majority of natural numbers are products of distinct prime numbers, and thus elements of S.

9. (Ross P3.51) Prostate cancer is the most common type of cancer found in males. As an indicator of whether a male has prostate cancer, doctors often perform a test that measures the level of the prostate-specific antigen (PSA) that is produced only by the prostate gland. Although PSA levels are indicative of cancer, the test is notoriously unreliable. Indeed, the probability that a noncancerous man will have an elevated PSA level is approximately 0.135, increasing to approximately 0.268 if the man does have cancer. If, on the basis of other factors, a physician is 70 percent certain that a male has prostate cancer, what is the conditional probability that he has the cancer given that

   (a) the test indicated an elevated PSA level?

   (b) the test did not indicate an elevated PSA level?

   Repeat the preceding calculation, this time assuming that the physician initially believes that there is a 30 percent chance that the man has prostate cancer.

   Solution. We will use Bayesian Inference.
   Let H_1 = the man has prostate cancer.
   Let H_2 = the man does not have prostate cancer.
   Then $S = H_1 \cup H_2$, and H_1 and H_2 are mutually disjoint.
   Let E = the test indicates an elevated PSA level. Then E^c = the test does not indicate an elevated PSA level.
   We are given that $\begin{cases} \mathbb{P}(E|H_1) = 0.268 \\ \mathbb{P}(E|H_2) = 0.135 \end{cases}$
   We can easily compute that $\begin{cases} \mathbb{P}(E^c|H_1) = 1 - 0.268 = 0.732 \\ \mathbb{P}(E^c|H_2) = 1 - 0.135 = 0.865 \end{cases}$
   Physician initially believes there is a 70% chance that a male has prostate cancer: In this case, our best guesses at $\mathbb{P}(H_1)$ and $\mathbb{P}(H_2)$ are $\begin{cases} \mathbb{P}(H_1) = 0.70 \\ \mathbb{P}(H_2) = 1 - 0.70 = 0.30 \end{cases}$

   (a) We want to find
   $$\mathbb{P}(H_1|E)$$
   By Bayes' Theorem, we know
   $$\mathbb{P}(H_1|E) = \frac{\mathbb{P}(E|H_1)\mathbb{P}(H_1)}{\mathbb{P}(E|H_1)\mathbb{P}(H_1) + \mathbb{P}(E|H_2)\mathbb{P}(H_2)}$$
   We already know all of these values, so we can plug them in to calculate that
   $$\mathbb{P}(H_1|E) = \frac{0.268 \cdot 0.70}{0.268 \cdot 0.70 + 0.135 \cdot 0.30} = \frac{0.1876}{0.2281} \approx 82.24\%$$
   Thus, the conditional probability that the male has cancer given that the test indicated an elevated PSA level is
   $$\mathbb{P}(H_1|E) \approx 82.24\%$$

   (b) This time, we want to find
   $$\mathbb{P}(H_1|E^c)$$
   By Bayes' Theorem, we know
   $$\mathbb{P}(H_1|E^c) = \frac{\mathbb{P}(E^c|H_1)\mathbb{P}(H_1)}{\mathbb{P}(E^c|H_1)\mathbb{P}(H_1) + \mathbb{P}(E^c|H_2)\mathbb{P}(H_2)}$$
   We already know all of these values, so we can plug them in to calculate that
   $$\mathbb{P}(H_1|E^c) = \frac{0.732 \cdot 0.7}{0.732 \cdot 0.7 + 0.865 \cdot 0.3} = \frac{0.5124}{0.7719} \approx 66.38\%$$

Thus, the probability that the male has prostate cancer given the test did not indicate elevated PSA levels is $\mathbb{P}(H_1|E^c) \approx 66.38\%$.

Physician believes there is a 30% chance that a male has prostate cancer:

In this case, our best guesses at $\mathbb{P}(H_1)$ and $\mathbb{P}(H_2)$ are $\begin{cases} \mathbb{P}(H_1) = 0.30 \\ \mathbb{P}(H_2) = 1 - 0.30 = 0.70 \end{cases}$

The other given probabilities remain the same as in the previous calculations.

(a) Once again, we want to find

$$\mathbb{P}(H_1|E)$$

Bayes' Theorem guarantees that

$$\mathbb{P}(H_1|E) = \frac{\mathbb{P}(E|H_1)\mathbb{P}(H_1)}{\mathbb{P}(E|H_1)\mathbb{P}(H_1) + \mathbb{P}(E|H_2)\mathbb{P}(H_2)}$$

We already know all of these values, so we can plug them in to calculate that

$$\mathbb{P}(H_1|E) = \frac{0.268 \cdot 0.30}{0.268 \cdot 0.30 + 0.135 \cdot 0.7} = \frac{0.0804}{0.1749} \approx 45.97\%$$

Thus, the probability that the male has prostate cancer given that the test did indicate elevated PSA levels is $\mathbb{P}(H_1|E) \approx 45.97\%$.

(b) Once again, we want to find

$$\mathbb{P}(H_1|E^c)$$

By Bayes' Theorem, we know

$$\mathbb{P}(H_1|E^c) = \frac{\mathbb{P}(E^c|H_1)\mathbb{P}(H_1)}{\mathbb{P}(E^c|H_1)\mathbb{P}(H_1) + \mathbb{P}(E^c|H_2)\mathbb{P}(H_2)}$$

We already know all of these values, so we can plug them in to calculate that

$$\mathbb{P}(H_1|E^c) = \frac{0.732 \cdot 0.30}{0.732 \cdot 0.30 + 0.865 \cdot 0.7} = \frac{0.2196}{0.8251} \approx 26.61\%$$

Thus, the probability that the male has prostate cancer given that the test did not indicate elevated PSA levels is $\mathbb{P}(H_1|E^c) \approx 26.61\%$.

Assignment 7

Math 407 (Swanson) – Spring 2023
Homework 1
Due Friday 1/13, 11:59pm

Name: Emerson Kahle Section: 39981

- You must upload your solutions to Gradescope as **one single, high-quality PDF**. You can convert paper-based work to a high-quality PDF using a scanning app for mobile devices, such as Adobe Scan (free, available for iOS and Android, can do multiple pages) or many others. If necessary, you can combine or merge multiple PDF's into a single PDF using a variety of services, such as Adobe Acrobat's cloud-based merge tool.
- After you upload, you must match each question with its corresponding page using Gradescope's interface. This allows graders to spend more time giving you feedback instead of hunting through submissions.
- Answers without supporting work will receive no credit. Show your work.
- You are encouraged to work together on homework, but **you must write up your solutions separately in your own words**. Copying from your fellow students or other sources is a serious academic integrity violation. In particular, you may not use "tutoring" services which simply provide answers.
- You are encouraged to typeset your solutions in LaTeX. Source code has been provided on Blackboard. Overleaf is a popular cloud-based editor.
- Problem numbers refer to the course textbook, though the problems may have been modified significantly.

1. (Ross P3.7) Suppose the king comes from a family of 2 children.

   (a) Under agnatic primogeniture, the oldest male child of a monarch inherits the throne. In such a system, what is the probability that the other child is the king's sister?

   (b) Under absolute primogeniture, the oldest child of a monarch, regardless of gender, inherits the throne. In such a system, what is the probability that the other child is the king's sister?

   Solution.

   (a) Consider the sample space, S, which is the set of all possible sequences of genders for the two children. Then

   $$S = \{(boy, boy), \quad (boy, girl), \quad (girl, boy), \quad (girl, girl)\}$$

   where all outcomes are equally likely. Since we know a king comes from the two children, we know that at least one of the two children is a boy. Since the oldest male child becomes king, we know that under agnatic primogeniture, if at least one of the two children is a boy, one of the two children will become king. Therefore, we know there is a king from a family of 2 children under agnatic primogeniture $\iff$ at least one of the two children is a boy. Let B = at least one of the two children is a boy. Then

   $$B = \{(boy, boy), \quad (boy, girl), \quad (girl, boy)\}$$

   We want to find the probability, given that at least one of the children is a boy, the other child is a girl (and thus the king's sister). Let G = one of the two children is a boy and the other is a girl. Then, if we are given that the king comes from one of the two children, we know that the other child will be the king's sister $\iff$ one child is a boy and the other is a girl. Therefore, the probability that the other child is the king's sister given that the king comes from a family of 2 children (under agnatic primogeniture) is $\mathbb{P}(G|B)$. We have

   $$G = \{(boy, girl), \quad (girl, boy)\}$$

   By the definition of conditional probability, we know

   $$\mathbb{P}(G|B) = \frac{\mathbb{P}(GB)}{\mathbb{P}(B)} \qquad (1)$$

   Comparing G and B, we see that they only share $(girl, boy)$ and $(boy, girl)$, so

   $$GB = \{(girl, boy), \quad (boy, girl)\}$$

   Since all outcomes in S are equally likely, we know that

   $$\mathbb{P}(GB) = \frac{|GB|}{|S|} = \frac{2}{4} = \frac{1}{2} = 50\%$$

   Similarly, since all outcomes in S are equally likely, we know that

   $$\mathbb{P}(B) = \frac{|B|}{|S|} = \frac{3}{4} = 75\%$$

   Plugging these values into (1), we find that the probability (under agnatic primogeniture) that the other child is the king's sister given that the king comes from a family of two children is

   $$\mathbb{P}(G|B) = \frac{\mathbb{P}(GB)}{\mathbb{P}(B)} = \frac{\frac{2}{4}}{\frac{3}{4}} = \frac{2}{3} \approx 66.67\%$$

(b) Our sample space, S, remains unchanged from part (a), and all outcomes are still equally likely. However, since the oldest child, regardless of gender inherits the throne under absolute primogeniture, the king will come from a family of two children $\iff$ the oldest of those two children is a boy. Let B' = the even that the oldest of the two children is a boy. Then

$$B' = \{(boy, boy), \quad (boy, girl)\}$$

and, since all outcomes are equally likely, we know

$$\mathbb{P}(B') = \frac{|B'|}{|S|} = \frac{2}{4} = \frac{1}{2} = 50\%$$

Given that the oldest child is a boy (and thus the king comes from the two children under absolute primogeniture), we want to find the probability that the other child is a girl (and thus the king's sister). Let G' = the event that the oldest child is a boy and the other child is a girl. Then the probability that the other child is the king's sister given that the king comes from a family of two children under absolute primogeniture is

$$\mathbb{P}(G'|B') = \frac{\mathbb{P}(G'B')}{\mathbb{P}(B')} \qquad (2)$$

and we have

$$G' = \{(boy, girl)\}$$

Comparing G' and B', we see that they only share $(boy, girl)$, so we know

$$G'B' = \{(boy, girl)\}$$

Since all outcomes are equally likely, we know

$$\mathbb{P}(G'B') = \frac{|G'B'|}{|S|} = \frac{1}{4} = 25\%$$

Plugging $\mathbb{P}(B')$ and $\mathbb{P}(G'B')$ into (2), we find that the probability that the other child is the king's sister, given that the king comes from a family of two children under absolute primogeniture, is

$$\mathbb{P}(G'|B') = \frac{\mathbb{P}(G'B')}{\mathbb{P}(B')} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{2}{4} = \frac{1}{2} = 50\%$$

2. (Ross P3.12) Suppose distinct values are written on each of 3 cards, which are then randomly given the designations A, B, and C. Given that card A's value is less than card B's value, find the probability it is also less than card C's value.

Solution.
Let c denote a card and $v(c)$ denote that cards value. Let's order the values of the cards before assigning the designations A, B, and C. Since the values are all distinct, we know that we can always order the cards c_1, c_2, c_3 s.t. $v(c_1) < v(c_2) < v(c_3)$. There are $3! = 6$ possible ways to permute the designations A, B, and C among these three cards c_1, c_2, c_3 such that each card receives exactly one designation, and all designations are assigned to exactly one card. The list of such possible permutations forms our sample space, S. We have

$$\begin{aligned} S = \quad & \{(v(A=c_1) < v(B=c_2) < v(C=c_3)), \quad (v(A=c_1) < v(C=c_2) < v(B=c_3)), \\ & (v(B=c_1) < v(A=c_2) < v(C=c_3)), \quad (v(B=c_1) < v(C=c_2) < v(A=c_3)), \\ & (v(C=c_1) < v(A=c_2) < v(B=c_3)), \quad (v(C=c_1) < v(B=c_2) < v(A=c_3))\} \end{aligned}$$

where all outcomes are equally likely since the designations are assigned randomly.
Let A_B = the event that card A's value is less than card B's value.
Let A_C = the event that card A's value is less than card C's value.
Then

$$\begin{aligned} A_B = \quad & \{(v(A=c_1) < v(B=c_2) < v(C=c_3)), \quad (v(A=c_1) < v(C=c_2) < v(B=c_3)), \\ & (v(C=c_1) < v(A=c_2) < v(B=c_3))\} \end{aligned}$$

and

$$\begin{aligned} A_C = \quad & \{(v(A=c_1) < v(B=c_2) < v(C=c_3)), \quad (v(A=c_1) < v(C=c_2) < v(B=c_3)), \\ & (v(B=c_1) < v(A=c_2) < v(C=c_3))\} \end{aligned}$$

We want to find the probability that $v(A) < v(C)$ given that $v(A) < v(B)$, so we need to compute

$$\mathbb{P}(A_C|A_B) = \frac{\mathbb{P}(A_C A_B)}{\mathbb{P}(A_B)} \qquad (1)$$

Since all outcomes are equally likely, we know that

$$\mathbb{P}(A_B) = \frac{|A_B|}{|S|} = \frac{3}{6} = \frac{1}{2} = 50\%$$

Comparing A_B and A_C, we see they only share $(v(A=c_1) < v(B=c_2) < v(C=c_3))$ and $(v(A=c_1) < v(C=c_2) < v(B=c_3))$, so we know

$$A_C A_B = \{(v(A=c_1) < v(B=c_2) < v(C=c_3)), \quad (v(A=c_1) < v(C=c_2) < v(B=c_3))\}$$

Since all outcomes are equally likely, this implies that

$$\mathbb{P}(A_C A_B) = \frac{|A_C A_B|}{|S|} = \frac{2}{6} = \frac{1}{3} \approx 33.33\%$$

Plugging $\mathbb{P}(A_B)$ and $\mathbb{P}(A_C A_B)$ into (1), we find that the probability that card A's value is less than card C's value, given that it is less than card B's value, is

$$P(A_C|A_B) = \frac{\mathbb{P}(A_C A_B)}{\mathbb{P}(A_B)} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3} \approx 66.67\%$$

3. (Ross P3.14) There are 15 tennis balls in a box, of which 9 have not previously been used. Three of the balls are randomly chosen, played with, and then returned to the box. Later, another 3 balls are randomly chosen from the box. Find the probability that none of these balls has ever been used.

Solution.
Let U_i = the event that exactly i of the three tennis balls we first randomly select are previously used. Since there are $6 > 3$ previously used balls and $9 > 3$ new balls, we could first select 0 previously used balls, or 1 previously used ball, or 2 previously used balls, or 3 previously used balls. Since we are only first selecting three balls, these are the only possible numbers of previously used balls which we could select. Therefore, we know that $\mathbb{P}(U_i) \begin{cases} \neq 0 \text{ if } 0 \leq i \leq 3 \\ = 0 \text{ otherwise} \end{cases}$ and $S = \bigcup_{i=0}^{3} U_i$.
We want to compute the probability that *none* of the three balls later chosen from the box has been previously used.
Let N = the event that none of the three balls later chosen from the box has been previously used.
Since for all $u_i \in U_i$, $u_i \notin U_j$ for all $0 \leq j \neq i \leq 3$, we know the U_i's are mutually disjoint. Combining this with the fact that $S = \bigcup_{i=0}^{3} U_i$, we can apply the Law of Total Probability to conclude that the probability that *none* of the three balls later chosen from the box has been previously used is

$$\mathbb{P}(N) = \quad \mathbb{P}(N|U_0)\mathbb{P}(U_0) + \mathbb{P}(N|U_1)\mathbb{P}(U_1) + \mathbb{P}(N|U_2)\mathbb{P}(U_2) + \mathbb{P}(N|U_3)\mathbb{P}(U_3) \qquad (1)$$

First, let's compute $\mathbb{P}(U_i)$ for all $0 \leq i \leq 3$.
For all such i, since we are first selecting three balls randomly from a set of 15, the sample space, S, has

$$|S| = \binom{15}{3}$$

equally likely outcomes.

(i) $\mathbb{P}(U_0)$: Since we first select exactly 0 previously used balls, we know all three of the balls we first select are new. There are only 9 new balls, which gives us

$$|U_0| = \binom{9}{3}$$

equally likely ways to do this. Since all outcomes are equally likely, the probability we first select exactly 0 previously used balls is

$$\mathbb{P}(U_0) = \frac{|U_0|}{|S|} = \frac{\binom{9}{3}}{\binom{15}{3}} \approx 18.46\%$$

(ii) $\mathbb{P}(U_1)$: Since we first select exactly 1 previously used ball, we know exactly 2 of the three balls we first select are new. There are 6 previously used balls, from which we must select exactly 1. There are $\binom{6}{1}$ equally likely ways to do this. For each of these selections, we must choose 2 of the 9 new balls, which can be done in $\binom{9}{2}$ ways. Thus, we know that U_1 has

$$|U_1| = \binom{6}{1}\binom{9}{2}$$

elements. Since all outcomes are equally likely, the probability that we first select exactly 1 previously used ball is

$$\mathbb{P}(U_1) = \frac{|U_1|}{|S|} = \frac{\binom{6}{1}\binom{9}{2}}{\binom{15}{3}} \approx 47.47\%$$

(iii) $\mathbb{P}(U_2)$: Since we first select exactly 2 previously used balls, we know exactly 1 of the three balls we first select is new. There are 6 previously used balls, from which we must select exactly 2. There

are $\binom{6}{2}$ equally likely ways to do this. For each of these combinations of 2 previously used balls, we must choose 1 of the 9 new balls, which can be done in $\binom{9}{1}$ ways. Thus, we know that U_2 has

$$|U_2| = \binom{6}{2}\binom{9}{1}$$

elements. Since all outcomes are equally likely, the probability that we first select exactly 2 previously used balls is

$$\mathbb{P}(U_2) = \frac{|U_2|}{|S|} = \frac{\binom{6}{2}\binom{9}{1}}{\binom{15}{3}} \approx 29.67\%$$

(iv) $\mathbb{P}(U_3)$: Since we first select exactly 3 previously used balls, we know none of the three balls we first select is new. There are 6 previously used balls, from which we must select exactly 3. There are $|U_3| = \binom{6}{3}$ ways to do this. Since all outcomes are equally likely, the probability that we first select exactly 3 previously used balls is

$$\mathbb{P}(U_3) = \frac{|U_3|}{|S|} = \frac{\binom{6}{3}}{\binom{15}{3}} \approx 4.40\%$$

Now, lets calculate $\mathbb{P}(N|U_i)$ for all $0 \leq i \leq 3$. Since we are still just selecting three balls from a set of 15 balls, the size of the sample space is still $|S| = \binom{15}{3}$ as defined above.

(i) $\mathbb{P}(N|U_0)$: Since none of the first three balls we select is previously used, we first select three new balls, so when we later select three balls, $6 + 3 = 9$ of them are previously used, and $9 - 3 = 6$ of them are new. At this point, to select no previously used balls, we need to later select 3 new balls from the 6 remaining new balls, which can be done in $\binom{6}{3}$ ways. Since there are $|S| = \binom{15}{3}$ ways to later select the three balls, the probability that we later select no previously used balls given that we first select no previously used balls is

$$\mathbb{P}(N|U_0) = \frac{\binom{6}{3}}{\binom{15}{3}} \approx 4.40\%$$

(ii) $\mathbb{P}(N|U_1)$: Since exactly 1 of the first three balls we select is previously used, we first select exactly 2 new balls. When we later select three balls, $6 + 2 = 8$ of them are previously used and $9 - 2 = 7$ of them are new. At this point, to select no previously used balls, we need to later select 3 balls from the 7 remaining new balls, which can be done in $\binom{7}{3}$ ways. Since there are $|S| = \binom{15}{3}$ ways to later select the three balls, the probability that we later select no previously used balls given that we first select exactly 1 previously used ball is

$$\mathbb{P}(N|U_1) = \frac{\binom{7}{3}}{\binom{15}{3}} \approx 7.69\%$$

(iii) $\mathbb{P}(N|U_2)$: Since exactly 2 of the first three balls we select are previously used, we first select exactly 1 new ball. When we later select three balls, $6 + 1 = 7$ of them are previously used and $9 - 1 = 8$ of them are new. At this point, to select no previously used balls, we need to later select 3 balls from the 8 remaining new balls, which can be done in $\binom{8}{3}$ ways. Since there are $|S| = \binom{15}{3}$ ways to later select the three balls, the probability that we later select no previously used balls given that we first select exactly 1 previously used ball is

$$\mathbb{P}(N|U_2) = \frac{\binom{8}{3}}{\binom{15}{3}} \approx 12.31\%$$

(iv) $\mathbb{P}(N|U_3)$: Since exactly 3 of the first three balls we select is previously used, we first select exactly 0 new balls. When we later select three balls, $6 + 0 = 6$ of them are previously used and $9 - 0 = 0$

of them are new. At this point, to select no previously used balls, we need to later select 3 balls from the 9 remaining new balls, which can be done in $\binom{9}{3}$ ways. Since there are $|S| = \binom{15}{3}$ ways to later select the three balls, the probability that we later select no previously used balls given that we first select exactly 1 previously used ball is

$$\mathbb{P}(N|U_3) = \frac{\binom{9}{3}}{\binom{15}{3}} \approx 18.46\%$$

Plugging all of these probabilities into (1), we find the probability that none of the three balls later chosen from the box has been previously used is

$$\begin{aligned}\mathbb{P}(N) = &\quad \mathbb{P}(N|U_0)\mathbb{P}(U_0) + \mathbb{P}(N|U_1)\mathbb{P}(U_1) + \mathbb{P}(N|U_2)\mathbb{P}(U_2) + \mathbb{P}(N|U_3)\mathbb{P}(U_3) \\ = &\quad \frac{\binom{6}{3}}{\binom{15}{3}}\frac{\binom{9}{3}}{\binom{15}{3}} + \frac{\binom{7}{3}}{\binom{15}{3}}\frac{\binom{6}{1}\binom{9}{2}}{\binom{15}{3}} + \frac{\binom{8}{3}}{\binom{15}{3}}\frac{\binom{6}{2}\binom{9}{1}}{\binom{15}{3}} + \frac{\binom{9}{3}}{\binom{15}{3}}\frac{\binom{6}{3}}{\binom{15}{3}} \approx 8.93\%\end{aligned}$$

4. (Ross P3.39)

   (a) A gambler has a fair coin and a two-headed coin in his pocket. He selects one of the coins at random; when he flips it, it shows heads. What is the probability that it is the fair coin?

   (b) Suppose that he flips the same coin a second time and, again, it shows heads. Now what is the probability that it is the fair coin?

   (c) Suppose that he flips the same coin a third time and it shows tails. Now what is the probability that it is the fair coin?

Solution.

(a) Let F = the event that the gambler selects the fair coin.
Let H = the event that the gambler selects the two-headed coin.
Since the gambler selects one of these two coins at random, we know that $\mathbb{P}(F) = \mathbb{P}(H)$. Also, since the gambler selects only one of the coins, we know H and F are mutually disjoint. Since the gambler must either select the fair or the two-headed coin, we know that $S = F \cup H$.
Let A = the even that the gambler's randomly selected coin lands on heads.
We want to find $\mathbb{P}(F|A)$. Applying Bayes' Theorem, we find

$$\mathbb{P}(F|A) = \frac{\mathbb{P}(A|F)\mathbb{P}(F)}{\mathbb{P}(A)} \qquad (1)$$

We can quickly compute the two probabilities in the numerator.
For $\mathbb{P}(A|F)$, we know that the gambler selected the fair coin, so there is a 50% chance of heads and a 50% chance of tails. Thus, the probability that the gambler's randomly selected coin lands on heads, given that the gambler selected the fair coin is

$$\mathbb{P}(A|F) = \frac{1}{2} = 50\%$$

For $\mathbb{P}(F)$, since $\mathbb{P}(F) = \mathbb{P}(H)$, F and H are mutually disjoint, and $S = H \cup F$, we know that

$$\mathbb{P}(S) = 1 = \mathbb{P}(H) + \mathbb{P}(F) = 2\mathbb{P}(F) \implies \mathbb{P}(F) = \frac{1}{2} = 50\%$$

We can also use the fact that H and F are mutually disjoint events whose union contains the entire sample space to calculate $\mathbb{P}(A)$. This allows us to apply the Law of Total Probability and find that

$$\mathbb{P}(A) = \mathbb{P}(A|F)\mathbb{P}(F) + \mathbb{P}(A|H)\mathbb{P}(H) \qquad (2)$$

We already found $\mathbb{P}(A|F)$ and $\mathbb{P}(F)$.
For $\mathbb{P}(A|H)$, we know that the gambler selected the two-headed coin, so there is a 100% chance that the coin lands on heads. Thus the probability that the gambler's randomly selected coin lands on heads, given that the gambler selected the two-headed coin is

$$\mathbb{P}(A|H) = 1 = 100\%$$

For $\mathbb{P}(H)$, we know that $\mathbb{P}(F) = 50\%$ and $\mathbb{P}(F) = \mathbb{P}(H)$. This allows us to conclude that the probability that the gambler's randomly selected coin is the two-headed coin is

$$\mathbb{P}(H) = \frac{1}{2} = 50\%$$

Plugging these values into (2), we find that the probability that the gambler's randomly selected coin lands on heads is

$$\mathbb{P}(A) = \mathbb{P}(A|F)\mathbb{P}(F) + \mathbb{P}(A|H)\mathbb{P}(H) = \frac{1}{2}\frac{1}{2} + (1)\frac{1}{2} = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} = 75\%$$

Plugging this into (1), we find that the probability that the gambler selected the fair coin given that it shows heads is

$$\mathbb{P}(F|A) = \frac{\mathbb{P}(A|F)\mathbb{P}(F)}{\mathbb{P}(A)} = \frac{\frac{1}{2}\frac{1}{2}}{\frac{3}{4}} = \frac{1}{4}\frac{4}{3} = \frac{1}{3} \approx 33.33\%$$

(b) Let B = the event that the gambler's randomly selected coin lands on head twice in a row.
We want to find $\mathbb{P}(F|B)$. Applying Bayes' Theorem, we find

$$\mathbb{P}(F|B) = \frac{\mathbb{P}(B|F)\mathbb{P}(F)}{\mathbb{P}(B)} \qquad (3)$$

We already computed $\mathbb{P}(F) = \frac{1}{2} = 50\%$ and we can quickly compute the other probability in the numerator.
For $\mathbb{P}(B|F)$, we know that the gambler selected the fair coin, so there is a 50% chance of heads and a 50% chance of tails on each coin flip. Thus, the probability that the gambler's randomly selected coin lands on heads twice in a row, given that the gambler selected the fair coin is

$$\mathbb{P}(A|F) = \left(\frac{1}{2}\right)^2 = \frac{1}{2}\frac{1}{2} = \frac{1}{4} = 25\%$$

Similar to part (a), we can use the Law of Total Probability to calculate $\mathbb{P}(B)$, which tells us that the probability that the gambler's randomly selected coin lands on heads twice in a row is

$$\mathbb{P}(B) = \mathbb{P}(B|F)\mathbb{P}(F) + \mathbb{P}(B|H)\mathbb{P}(H) \qquad (4)$$

We already found $\mathbb{P}(B|F)$, $\mathbb{P}(F)$, and $\mathbb{P}(H)$.
For $\mathbb{P}(B|H)$, we know that the gambler selected the two-headed coin, so there is a 100% chance that the coin lands on heads on every coin toss. Thus, the probability that the gambler's randomly selected coin lands on heads twice in a row, given that the gambler selected the two-headed coin, is

$$\mathbb{P}(B|H) = 1^2 = 1 = 100\%$$

Plugging these values into (4), we find that the probability hat the gambler's randomly selected coin lands on heads twice in a row is

$$\mathbb{P}(B) = \mathbb{P}(B|F)\mathbb{P}(F) + \mathbb{P}(B|H)\mathbb{P}(H) = \frac{1}{4}\frac{1}{2} + (1)\frac{1}{2} = \frac{1}{8} + \frac{1}{2} = \frac{1}{8} + \frac{4}{8} = \frac{5}{8} = 62.5\%$$

Plugging this into (3), we find that the probability that the gambler selected the fair coin given that it lands on heads twice in a row is

$$\mathbb{P}(F|B) = \frac{\mathbb{P}(B|F)\mathbb{P}(F)}{\mathbb{P}(B)} = \frac{\frac{1}{4}\frac{1}{2}}{\frac{5}{8}} = \frac{1}{8}\frac{8}{5} = \frac{1}{5} = 20\%$$

(c) Let C = the event that the gambler's randomly selected coin lands on heads twice and then tails.
We want to find $\mathbb{P}(F|C)$. Applying Bayes' Theorem, we find

$$\mathbb{P}(F|C) = \frac{\mathbb{P}(C|F)\mathbb{P}(F)}{\mathbb{P}(C)} \qquad (5)$$

We already found $\mathbb{P}(F)$, and we can easily compute $\mathbb{P}(C|F)$. Since we know the gambler's randomly selected coin is the fair coin, we know there is a $\frac{1}{2} = 50\%$ chance of the coin landing on heads and a $\frac{1}{2} = 50\%$ chance of the coin landing on tails for all coin tosses. Therefore, the probability that the gambler's randomly selected coin lands on heads twice and then tails, given that the gambler selected the fair coin is

$$\mathbb{P}(C|F) = \left(\frac{1}{2}\right)^3 = \frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8} = 12.5\%$$

Similar to parts (a) and (b), we can apply the Law of Total Probability to calculate $\mathbb{P}(C)$, which tells us that the probability that the gambler's randomly selected coin lands on heads twice and then tails is

$$\mathbb{P}(C) = \mathbb{P}(C|F)\mathbb{P}(F) + \mathbb{P}(C|H)\mathbb{P}(H) \qquad (6)$$

We already computed $\mathbb{P}(C|F)$, $\mathbb{P}(F)$, and $\mathbb{P}(H)$, so we just need to find $\mathbb{P}(C|H)$. Since we know the gambler selected the two-headed coin, we know there is a 0% chance that the coin lands on tails on each toss. Thus, the probability that the gambler's randomly selected coin lands on heads twice and then tails is

$$\mathbb{P}(C|H) = 1^2 \cdot 0 = 0 = 0\%$$

Plugging these values into (6), we find that the probability that the gambler's randomly selected coin lands on heads twice and then tails is

$$\mathbb{P}(C) = \mathbb{P}(C|F)\mathbb{P}(F) + \mathbb{P}(C|H)\mathbb{P}(H) = \frac{1}{8}\frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{16}$$

Plugging this into (5), we find that the probability that the gambler's randomly selected coin is fair given that it lands on heads twice and then tails is

$$\mathbb{P}(F|C) = \frac{\mathbb{P}(C|F)\mathbb{P}(F)}{\mathbb{P}(C)} = \frac{\frac{1}{8}\frac{1}{2}}{\frac{1}{16}} = \frac{1}{16}\frac{16}{1} = 1 = 100\%$$

This result follows basic intuition, as it is impossible for the two-headed coin to ever land on tails. Thus, as soon as we know that the gambler's randomly selected coin lands on tails once, we know that it is the fair coin.

5. (Ross P3.46) Three prisoners are informed by their jailer that one of them has been chosen at random to be executed and the other two are to be freed. Prisoner A asks the jailer to tell him privately which of his fellow prisoners will be set free, claiming that there would be no harm in divulging this information because he already knows that at least one of the two will go free. The jailer refuses to answer the question, pointing out that if A knew which of his fellow prisoners were to be set free, then his own probability of being executed would rise from $\frac{1}{3}$ to $\frac{1}{2}$ because he would then be one of two prisoners. What do you think of the jailer's reasoning?

Solution.
Let E_A = the event that Prisoner A is to be executed.
Let E_B = the event that Prisoner B is to be executed.
Let E_C = the event that Prisoner C is to be executed.
Note that, since only one prisoner is to be executed, E_A, E_B, and E_C are all mutually disjoint. Since one of Prisoners A, B, and C is guaranteed to be executed, $S = E_A \cup E_B \cup E_C$. Let T_B = the event that Prisoner A is told that Prisoner B is to be executed.
Let T_C = the event that Prisoner A is told that Prisoner C is to be executed.
Then we can interpret the jailer's reasoning as

$$\mathbb{P}(E_A|T_B) = \mathbb{P}(E_A|T_C) = \frac{1}{2} > \frac{1}{3} = \mathbb{P}(E_A) \qquad (1)$$

Let's calculate these probabilities individually to see if we agree with the jailer.
For $\mathbb{P}(E_A)$, since there are three prisoners, one of whom is randomly selected to be executed, we find that

$$\mathbb{P}(E_A) = \mathbb{P}(E_B) = \mathbb{P}(E_C) = \frac{1}{3} \approx 33.33\%$$

For $\mathbb{P}(E_A|T_B)$, we can apply Bayes' Theorem to find that

$$\mathbb{P}(E_A|T_B) = \frac{\mathbb{P}(T_B|E_A)\mathbb{P}(E_A)}{\mathbb{P}(T_B)}$$

Since E_A, E_B, and E_C are mutually disjoint events whose union is the sample space, we can apply the Law of Total Probability to find that

$$\mathbb{P}(E_A|T_B) = \frac{\mathbb{P}(T_B|E_A)\mathbb{P}(E_A)}{\mathbb{P}(T_B)} = \frac{\mathbb{P}(T_B|E_A)\mathbb{P}(E_A)}{\mathbb{P}(T_B|E_A)\mathbb{P}(E_A) + \mathbb{P}(T_B|E_B)\mathbb{P}(E_B) + \mathbb{P}(T_B|E_C)\mathbb{P}(E_C)} \qquad (2)$$

We already computed $\mathbb{P}(E_A)$, $\mathbb{P}(E_B)$, and $\mathbb{P}(E_C)$, so we just need to compute $\mathbb{P}(T_B|E_A)$, $\mathbb{P}(T_B|E_B)$, and $\mathbb{P}(T_B|E_C)$.
For $\mathbb{P}(T_B|E_A)$, since Prisoner A is to be executed, the jailer has two options for which prisoner he tells Prisoner A is to be freed. One of these options is Prisoner B, so assuming the jailer randomly selects one of the two options, the probability that the jailer tells Prisoner A that Prisoner B is to be freed, given that Prisoner A is to be executed, is

$$\mathbb{P}(T_B|E_A) = \frac{1}{2} = 50\%$$

For $\mathbb{P}(T_B|E_B)$, since Prisoner B is to be executed, the jailer cannot tell Prisoner A that Prisoner B is to be freed, assuming the jailer is truthful. Therefore, the probability that the jailer tells Prisoner A that Prisoner B is to be freed, given that Prisoner B is to be executed, is

$$\mathbb{P}(T_B|E_B) = 0 = 0\%$$

For $\mathbb{P}(T_B|E_C)$, since Prisoner C is to be executed, the jailer cannot tell Prisoner A that Prisoner C is to be freed, so the jailer must tell Prisoner A that Prisoner B is to be freed. Thus, the probability that the jailer tells Prisoner A that Prisoner B is to be freed, given that Prisoner C is to be executed, is

$$\mathbb{P}(T_B|E_B) = 1 = 100\%$$

Plugging these values into (2), we find that the probability that Prisoner A is executed given that the jailer tells Prisoner A that Prisoner B is to be freed is

$$\begin{aligned}\mathbb{P}(E_A|T_B) = & \quad \frac{\mathbb{P}(T_B|E_A)\mathbb{P}(E_A)}{\mathbb{P}(T_B|E_A)\mathbb{P}(E_A) + \mathbb{P}(T_B|E_B)\mathbb{P}(E_B) + \mathbb{P}(T_B|E_C)\mathbb{P}(E_C)} \\ = & \quad \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}\frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{\frac{1}{6}}{\frac{1}{6} + \frac{1}{3}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{6}\frac{6}{3} = \frac{1}{3} \approx 33.33\%\end{aligned}$$

By the symmetry of the problem, we can apply the exact same process to find that the probability that Prisoner A is executed given that the jailer tells Prisoner A that Prisoner C is to be freed is

$$\begin{aligned}\mathbb{P}(E_A|T_C) = & \quad \frac{\mathbb{P}(T_C|E_A)\mathbb{P}(E_A)}{\mathbb{P}(T_C|E_A)\mathbb{P}(E_A) + \mathbb{P}(T_C|E_B)\mathbb{P}(E_B) + \mathbb{P}(T_C|E_C)\mathbb{P}(E_C)} \\ = & \quad \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}\frac{1}{3} + 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{\frac{1}{6}}{\frac{1}{6} + \frac{1}{3}} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{6}\frac{6}{3} = \frac{1}{3} \approx 33.33\%\end{aligned}$$

To summarize our results, we found that

$$\mathbb{P}(E_A) = \frac{1}{3} = \mathbb{P}(E_A|T_B) = \mathbb{P}(E_A|T_C) \qquad (3)$$

Therefore, the probability that Prisoner A is executed *does not* rise from $\frac{1}{3}$ to $\frac{1}{2}$ by the jailer telling Prisoner A that one of the other prisoners will be set free. In fact, it doesn't change at all. Comparing (1) and (3), we see that our results directly contradict the implications of the jailer's reasoning. Therefore, I do not agree with the jailer's reasoning.

6. (Ross P3.48) In any given year, a male automobile policyholder will make a claim with probability p_m and a female policyholder will make a claim with probability p_f, where $p_f \neq p_m$. The fraction of the policyholders that are male is α, where $0 < \alpha < 1$. A policyholder is randomly chosen. If A_i denotes the event that this policyholder will make a claim in year i, show that

$$P(A_2 \mid A_1) > P(A_1)$$

Give an intuitive explanation of why the preceding inequality is true.

Solution.
Let $M =$ the event that the randomly selected policyholder is male.
We are given that

$$\mathbb{P}(M) = \alpha$$

Let $F =$ the event that the randomly selected policyholder is female.
Since the randomly selected policyholder is either male or female, we know $S = F \cup M$. Therefore, we know

$$\mathbb{P}(S) = 1 = \mathbb{P}(M) + \mathbb{P}(F) = \alpha + \mathbb{P}(F) \implies \mathbb{P}(F) = 1 - \alpha$$

Also, since the policyholder cannot be both male and female, we know that F and M are mutually disjoint. This allows us to apply the Law of Total Probability to find that

$$\mathbb{P}(A_1) = \mathbb{P}(A_1|M)\mathbb{P}(M) + \mathbb{P}(A_1|F)\mathbb{P}(F)$$

We are given that $\mathbb{P}(A_1|M) = p_m$ and $\mathbb{P}(A_1|F) = p_f$, which allows us to directly compute that the probability that the randomly selected policyholder makes a claim in year 1 is

$$\mathbb{P}(A_1) = \mathbb{P}(A_1|M)\mathbb{P}(M) + \mathbb{P}(A_1|F)\mathbb{P}(F) = p_m \cdot \alpha + p_f \cdot (1 - \alpha)$$

For $\mathbb{P}(A_2|A_1)$, we can apply the definition of conditional probability to find that the probability that the randomly selected policyholder makes a claim in year 2, given that they made a claim in year 1, is

$$\mathbb{P}(A_2|A_1) = \frac{\mathbb{P}(A_2A_1)}{\mathbb{P}(A_1)} = \frac{\mathbb{P}(A_2A_1)}{p_m \cdot \alpha + p_f \cdot (1 - \alpha)} \qquad (1)$$

We can use the Law of Total Probability once again to find that the probability that the randomly selected policyholder makes a claim on year 1 and year 2 is

$$\mathbb{P}(A_2A_1) = \mathbb{P}(A_2A_1|M)\mathbb{P}(M) + \mathbb{P}(A_2A_1|F)\mathbb{P}(F)$$

For $\mathbb{P}(A_2A_1|M)$, we know that the policyholder is a male, so we know there is a p_m probability of him making a claim on each year, so the probability that he makes a claim on both year 1 and year 2 is

$$\mathbb{P}(A_2A_1|M) = p_m^2$$

Similarly, for $\mathbb{P}(A_2A_1|F)$, we know that the policyholder is a female, so we know there is a p_f probability of her making a claim on each year, so the probability that she makes a claim on both year 1 and year 2 is

$$\mathbb{P}(A_2A_1|F) = p_f^2$$

This allows us to directly compute that the probability that the randomly selected policyholder makes a claim on both year 1 and year 2 is

$$\mathbb{P}(A_2A_1) = \mathbb{P}(A_2A_1|M)\mathbb{P}(M) + \mathbb{P}(A_2A_1|F)\mathbb{P}(F) = p_m^2 \cdot \alpha + p_f^2 \cdot (1 - \alpha)$$

Plugging this into (1), we find that the probability that the randomly selected policyholder makes a claim in year 2, given that they made a claim in year 1, is

$$\mathbb{P}(A_2|A_1) = \frac{\mathbb{P}(A_2A_1)}{p_m \cdot \alpha + p_f \cdot (1 - \alpha)} = \frac{p_m^2 \cdot \alpha + p_f^2 \cdot (1 - \alpha)}{p_m \cdot \alpha + p_f \cdot (1 - \alpha)}$$

Now that we have computed $\mathbb{P}(A_1)$ and $\mathbb{P}(A_2|A_1)$, we want to show that

$$\mathbb{P}(A_2|A_1) = \frac{p_m^2 \cdot \alpha + p_f^2 \cdot (1-\alpha)}{p_m \cdot \alpha + p_f \cdot (1-\alpha)} > p_m \cdot \alpha + p_f \cdot (1-\alpha) = \mathbb{P}(A_1)$$

It suffices to show that

$$p_m^2 \cdot \alpha + p_f^2 \cdot (1-\alpha) > \big(p_m \cdot \alpha + p_f \cdot (1-\alpha)\big)^2$$

so it suffices to show

$$p_m^2 \cdot \alpha + p_f^2 \cdot (1-\alpha) - \big(p_m \cdot \alpha + p_f \cdot (1-\alpha)\big)^2 > 0 \qquad (2)$$

Expanding and simplifying (2), we find that

$$\begin{aligned}
p_m^2 \cdot \alpha + p_f^2 \cdot (1-\alpha) - \big(p_m \cdot \alpha + p_f \cdot (1-\alpha)\big)^2 = \;& p_m^2 \cdot \alpha + p_f^2 - p_f^2 \cdot \alpha - 2p_m p_f \cdot \alpha \cdot (1-\alpha) - p_m^2 \cdot \alpha^2 \\
& - p_f^2 + 2\alpha \cdot p_f^2 - p_f^2 \cdot \alpha^2 \\
= \;& \alpha(p_m^2 + p_f^2 - p_m^2 \cdot \alpha - p_f^2 \cdot \alpha) - 2p_m p_f \cdot \alpha \cdot (1-\alpha) \\
= \;& \alpha(1-\alpha)(p_m^2 + p_f^2) - 2p_m p_f \alpha(1-\alpha) \\
= \;& \alpha(1-\alpha)(p_m^2 - 2p_m p_f + p_f^2) \\
= \;& \alpha(1-\alpha)(p_m - p_f)^2
\end{aligned}$$

Since $0 < \alpha < 1$, we know $0 < 1 - \alpha < 1$, so $\alpha(1-\alpha) > 0$.
Also, since $p_m \neq p_f$, we know that $p_m - p_f \neq 0 \implies (p_m - p_f)^2 > 0$.
This implies that

$$\alpha(1-\alpha)(p_m - p_f)^2 > 0$$

Therefore, we know that

$$p_m^2 \cdot \alpha + p_f^2 \cdot (1-\alpha) - \big(p_m \cdot \alpha + p_f \cdot (1-\alpha)\big)^2 = \alpha(1-\alpha)(p_m - p_f)^2 > 0$$

which completes the proof that

$$\mathbb{P}(A_2|A_1) = \frac{p_m^2 \cdot \alpha + p_f^2 \cdot (1-\alpha)}{p_m \cdot \alpha + p_f \cdot (1-\alpha)} > p_m \cdot \alpha + p_f \cdot (1-\alpha) = \mathbb{P}(A_1)$$

The intuition behind the solution relies on the facts that $p_m \neq p_f$ and $0 < \alpha < 1$. Since $0 < \alpha < 1$, we know that at least one policyholder is male and at least one policyholder is female. Since $p_m \neq p_f$, we know either $p_m > p_f$ or $p_f > p_m$.
If $p_m > p_f$, then male policyholders are more likely to make claims in a given year than female policyholders. Therefore, if we know the randomly selected policyholder makes a claim in year 1, we can intuit that they are more likely to be a man than they would be if we did not know that information. Since we know the policyholder is more likely to be a man since they made a claim in year 1, we can intuit they are more likely to make a claim in year 2 since $p_m > p_f$.
Similarly, if $p_f > p_m$, then female policyholders are more likely to make claims in a given year than male policyholders. Therefore, if we know the randomly selected policyholder makes a claim in year 1, we can intuit that they are more likely to be a woman than they would be if we did not know that information. Since we know the policyholder is more likely to be a woman since they made a claim in year 1, we can intuit they are more likely to make a claim in year 2 since $p_f > p_m$.
This is the intuition behind the result

$$\mathbb{P}(A_2|A_1) > P(A_1)$$

7. (Ross P3.57) In a 7 game series played with two teams, the first team to win a total of 4 games is the winner. Suppose that each game played is independently won by team A with probability p.

   (a) Given that one team leads 3 to 0, what is the probability that it is team A that leads?

   (b) Given that one team leads 3 to 0, what is the probability that team wins the series?

   Solution.

   (a) Let E_A = the event that team A leads 3 to 0.
   Call the other team team B, and let E_B = the event that team B leads 3 to 0.
   Let E_3 = the event that one team leads 3 to 0.
   Since team A and team B are the only two teams, $E_3 = E_A \cup E_B$.
   We want to find $\mathbb{P}(E_A|E_3)$. Applying the definition of conditional probability, we find the probability that team A leads 3 to 0 given one team leads 3 to 0 is

   $$\mathbb{P}(E_A|E_3) = \frac{\mathbb{P}(E_A E_3)}{\mathbb{P}(E_3)} \qquad (1)$$

   Since team A cannot lead 3 to 0 if team B leads 3 to 0, and vice versa, we know that E_A and E_B are mutually disjoint, so

   $$E_A E_3 = E_A(E_A \cup E_B) = E_A \cup E_A E_B = E_A \implies \mathbb{P}(E_A E_3) = \mathbb{P}(E_A)$$

   Now, we just need to compute $\mathbb{P}(E_A)$ and $\mathbb{P}(E_3)$.
   For $\mathbb{P}(E_A)$, we note that there is an independent probability of p that team A wins the game for each of the first three games. Therefore, the probability that team A wins all of the first three games (and thus leads 3 to 0) is

   $$\mathbb{P}(E_A) = p^3$$

   For $\mathbb{P}(E_3)$, we note that $\mathbb{P}(E_3) = \mathbb{P}(E_A \cup E_B) = \mathbb{P}(E_A) + \mathbb{P}(E_B)$ since E_A and E_B are mutually disjoint. We already know $\mathbb{P}(E_A) = p^3$, so we just need to calculate $\mathbb{P}(E_B)$. The probability that team A wins a given game is independently p for each game. Since team B must win for team A to lose, the probability that team B wins a given game is independently $1 - p$ for each game. Therefore, the probability that team B wins all of the first three games (and thus leads 3 to 0) is

   $$\mathbb{P}(E_B) = (1-p)^3$$

   Therefore, the total probability that one team leads 3 to 0 is

   $$\mathbb{P}(E_3) = \mathbb{P}(E_A) + \mathbb{P}(E_B) = p^3 + (1-p)^3$$

   Plugging $\mathbb{P}(E_A)$ and $\mathbb{P}(E_3)$ into (1), we find that the probability that team A leads 3 to 0 given that one team leads 3 to 0 is

   $$\mathbb{P}(E_A|E_3) = \frac{\mathbb{P}(E_A E_3)}{\mathbb{P}(E_3)} = \frac{\mathbb{P}(E_A)}{\mathbb{P}(E_3)} = \frac{p^3}{p^3 + (1-p)^3}$$

   (b) If one team leads 3 to 0 then wins the series, it could either be team A or team B.
   Let W_A = team A leads 3 to 0 then wins the series.
   Let W_B = team B leads 3 to 0 then wins the series.
   If team A leads 3 to 0 then wins the series, team B cannot lead 3 to 0 then win the series, so we know W_A and W_B are mutually disjoint.
   Let W_3 = one team leads 3 to 0 then wins the series. Then

   $$W_3 = W_A \cup W_B$$

We want to find $\mathbb{P}(W_3|E_3)$. Applying the definition of conditional probability, we find that the probability that one team leads 3 to 0 then wins the series, given that one team leads 3 to 0, is

$$\mathbb{P}(W_3|E_3) = \frac{\mathbb{P}(W_3E_3)}{\mathbb{P}(E_3)}$$

For all $w \in W_3$, one team leads 3 to 0 after three games, so $w \in E_3$. Therefore, we know that $W_3E_3 = W_3$, so we just need to find

$$\mathbb{P}(W_3|E_3) = \frac{\mathbb{P}(W_3)}{\mathbb{P}(E_3)} \qquad (2)$$

In part (a), we found that the probability that one team leads 3 to 0 after three games is

$$\mathbb{P}(E_3) = p^3 + (1-p)^3$$

so we just need to find $\mathbb{P}(W_3)$.
Since W_A and W_B are mutually disjoint, we know

$$\mathbb{P}(W_3) = \mathbb{P}(W_A \cup W_B) = \mathbb{P}(W_A) + \mathbb{P}(W_B) \qquad (3)$$

so we need to find $\mathbb{P}(W_A)$ and $\mathbb{P}(W_B)$.
For $\mathbb{P}(W_A)$, we know that team A needs to win the first 3 games, then Team B cannot win all 4 of the remaining games. The probability that team A wins the first 3 games is

$$\mathbb{P}(E_A) = p^3$$

as calculated in part (a). The probability that team B wins a given game is independently $1-p$, so the probability that team B wins all 4 of the remaining games is

$$(1-p)^4$$

Taking the complement, we find that the probability that team B *does not* win all 4 of the remaining games is

$$1-(1-p)^4$$

The results of the first 3 games are independent from the results of the last 4, so we can multiply these probabilities together to find that the total probability that team A leads 3 to 0 then wins the series is

$$\mathbb{P}(W_A) = p^3(1-(1-p)^4)$$

Similarly, for $\mathbb{P}(W_B)$, we know that team B needs to win the first 3 games, then team A cannot win all 4 of the remaining games. The probability that team B wins the first 3 games is

$$\mathbb{P}(E_B) = (1-p)^3$$

since there is an independent probability of $1-p$ that team B wins each game. The probability that team A wins a given game is independently p, so the probability that team A wins all 4 of the remaining games is

$$p^4$$

Taking the complement, we find that the probability that team A *does not* win all 4 of the remaining games is

$$1-p^4$$

The results of the first 3 games are independent from the results of the last 4, so we can multiply these probabilities together to find that the total probability that team B leads 3 to 0 then wins the series is

$$\mathbb{P}(W_B) = (1-p)^3(1-p^4)$$

Plugging $\mathbb{P}(W_B)$ and $\mathbb{P}(W_A)$ into (3), we find

$$\mathbb{P}(W_3) = \mathbb{P}(W_A) + \mathbb{P}(W_B) = p^3(1-(1-p)^4) + (1-p)^3(1-p^4)$$

Plugging this into (2), we find that, given that one team leads 3 to 0, the probability that team wins the series is

$$\mathbb{P}(W_3|E_3) = \frac{\mathbb{P}(W_3)}{\mathbb{P}(E_3)} = \frac{p^3(1-(1-p)^4) + (1-p)^3(1-p^4)}{p^3 + (1-p)^3}$$

8. (Ross P3.60) In a class, there are 4 first-year boys, 6 first-year girls, and 6 sophomore boys. How many sophomore girls must be present if gender and class are to be independent when a student is selected at random?

Solution.
Let $G =$ the event that the randomly selected student is a girl.
Let $H =$ the event that the randomly selected student is a sophomore.
Let $k =$ the number of sophomore girls in the class.
Then for gender and class to be independent when a student is selected at random, we need

$$\mathbb{P}(GH) = \mathbb{P}(G)\mathbb{P}(H) \qquad (1)$$

The sample space, S, is the set of all students in the class. There are 4 first-year boys, 6 first-year girls, 6 sophomore boys, and k sophomore girls, for a total of

$$|S| = 4 + 6 + 6 + k = 16 + k$$

students in the class. Since we are selecting one student randomly, there is an equal likelihood that each of these $16 + k$ students is selected.
Now, we can calculate each of the probabilities from (1) in terms of k, then plug them into (1) to solve for k.
For $\mathbb{P}(GH)$, note that $GH = \{$ all students that are girls and sophomores$\}$, and we know that exactly $|GH| = k$ of the students in the class are sophomore girls. Therefore, the probability that a randomly selected student is a sophomore and a girl is

$$\mathbb{P}(GH) = \frac{|GH|}{|S|} = \frac{k}{16 + k}$$

For $\mathbb{P}(G)$, we have k sophomore girls and 6 first-year girls, for a total of $|G| = k + 6$ girls in the class. Since we are equally likely to select each of these girls, the probability that a randomly selected student is a girl is

$$\mathbb{P}(G) = \frac{|G|}{|S|} = \frac{6 + k}{16 + k}$$

For $\mathbb{P}(H)$, we have k sophomore girls and 6 sophomore boys for a total of $|H| = k + 6$ sophomores in the class. Since we are equally likely to select each of these sophomores, the probability that a randomly selected student is a sophomore is

$$\mathbb{P}(H) = \frac{|H|}{|S|} = \frac{6 + k}{16 + k}$$

Plugging $\mathbb{P}(GH)$, $\mathbb{P}(G)$, and $\mathbb{P}(H)$ into (1), we find that

$$\frac{k}{16 + k} = \frac{6 + k}{16 + k} \cdot \frac{6 + k}{16 + k}$$

This is true $\iff$

$$k(16 + k)^2 = (16 + k)(6 + k)^2$$

This is true $\iff$

$$256k + 32k^2 + k^3 = 576 + 228k + 28k^2 + k^3$$

This is true $\iff$

$$4k^2 + 28k - 576 = 4(k^2 + 7k - 144) = 4(k + 16)(k - 9) = 0$$

This is true $\iff k = -16$ or $k = 9$.
Since we cannot have a negative number of sophomore girls in the class, this means we must have exactly $k = 9$ sophomore girls in the class for gender and class to be independent when a student is selected at

random.
We can verify that, when plugging in $k = 9$ into (1), the equality holds:

$$\mathbb{P}(GH) = \frac{9}{25} = \frac{3 \cdot 3}{25} = \frac{3 \cdot 3}{25}\frac{5 \cdot 5}{5 \cdot 5} = \frac{3 \cdot 5}{25}\frac{3 \cdot 5}{25} = \frac{15}{25}\frac{15}{25} = \mathbb{P}(G)\mathbb{P}(H)$$

as expected.

Assignment 8

Math 407 (Swanson) – Spring 2023
Homework 1
Due Friday 1/13, 11:59pm

Name: Emerson Kahle Section: 39981

- You must upload your solutions to Gradescope as **one single, high-quality PDF**. You can convert paper-based work to a high-quality PDF using a scanning app for mobile devices, such as Adobe Scan (free, available for iOS and Android, can do multiple pages) or many others. If necessary, you can combine or merge multiple PDF's into a single PDF using a variety of services, such as Adobe Acrobat's cloud-based merge tool.
- After you upload, you must match each question with its corresponding page using Gradescope's interface. This allows graders to spend more time giving you feedback instead of hunting through submissions.
- Answers without supporting work will receive no credit. Show your work.
- You are encouraged to work together on homework, but **you must write up your solutions separately in your own words**. Copying from your fellow students or other sources is a serious academic integrity violation. In particular, you may not use "tutoring" services which simply provide answers.
- You are encouraged to typeset your solutions in LaTeX. Source code has been provided on Blackboard. Overleaf is a popular cloud-based editor.
- Problem numbers refer to the course textbook, though the problems may have been modified significantly.

1. (Ross P3.63) Suppose that we want to generate the outcome of the flip of a fair coin, but that all we have at our disposal is a biased coin that lands on heads with some unknown probability p that need not equal to $\frac{1}{2}$. Consider the following procedure for accomplishing our task:

   1. Flip the coin.
   2. Flip the coin again.
   3. If both flips land on heads or both land on tails, return to step 1.
   4. Let the result of the last flip be the result of the experiment.

   (a) Show that the result is equally likely to be either heads or tails.

   (b) Could we use a simpler procedure that continues to flip the coin until the last two flips are different and then lets the result be the outcome of the final flip?

Solution.
Note: Under instruction from Professor Swanson, we assume that, although the coin may have $p \neq \frac{1}{2}$, the outcomes of any two flips are still independent, and $p \neq 0, 1$.

(a) Since we always flip the coin twice at one time, we can consider each pair of flips instead of individual flips.
Let TH = the event that a pair of flips is a tails followed by a heads.
Let HT = the event that a pair of flips is a heads followed by a tails.
HT and TH are mutually disjoint because once a pair is a heads followed by a tails, the result is tails, so there must not be any pair which is a tails followed by a heads, and vice versa.
In order for the experiment to have a result, we need to end on a pair which is either a heads followed by a tails (heads-tails) or a tails followed by a heads (tails-heads).
In order for that result to be heads, the last pair needs to be a (tails-heads) pair. Therefore, result of the experiment is heads $\iff$ we end on a (tails-heads) pair given that we get either a (tails-heads) pair or a (heads-tails) pair, so the probability that the result is heads is $\mathbb{P}(result = heads) = \mathbb{P}(TH|TH \cup HT)$.
Similarly, the result of the experiment is tails $\iff$ we end on a (heads-tails) pair given that we get either a (tails-heads) pair or a (heads-tails) pair, so the probability that the result is tails is $\mathbb{P}(result = tails) = \mathbb{P}(HT|TH \cup HT)$.
Now, we can calculate $\mathbb{P}(TH)$ and $\mathbb{P}(HT)$ to prove they are equal.

$\mathbb{P}(result = heads)$: Applying the definition of conditional probability, we find

$$\mathbb{P}(TH|TH \cup HT) = \frac{\mathbb{P}(TH \cap (TH \cup HT))}{\mathbb{P}(TH \cup HT)}$$

Since TH and HT are mutually disjoint, $\mathbb{P}(TH \cap (TH \cup HT)) = \mathbb{P}(TH \cap TH \cup TH \cap HT) = \mathbb{P}(TH)$ and $\mathbb{P}(TH \cup HT) = \mathbb{P}(TH) + \mathbb{P}(HT)$. This implies that

$$\mathbb{P}(TH|TH \cup HT) = \frac{\mathbb{P}(TH)}{\mathbb{P}(TH) + \mathbb{P}(HT)}$$

Since each coin flip is independent of the others,

$$\mathbb{P}(TH) = \mathbb{P}(T)\mathbb{P}(H) = (1-p)p = \mathbb{P}(H)\mathbb{P}(T) = \mathbb{P}(HT)$$

where T = the event that a given flip is tails and H = the event that a given flip is heads. Therefore,

$$\mathbb{P}(result = heads) = \mathbb{P}(TH|TH \cup HT) = \frac{(1-p)p}{(1-p)p + p(1-p)} = \frac{(1-p)p}{2(1-p)p} = \frac{1}{2} = 50\%$$

with the last step relying on $p \neq 0, 1$.

$\mathbb{P}(result = tails)$: Similarly applying the definition of conditional probability, we find

$$\mathbb{P}(HT|TH \cup HT) = \frac{\mathbb{P}(HT \cap (TH \cup HT))}{\mathbb{P}(TH \cup HT)}$$

Using the same implications of the independence of coin flips, we find

$$\begin{aligned}\mathbb{P}(result = tails) = \quad & \mathbb{P}(HT|TH \cup HT) = \frac{\mathbb{P}(HT)}{\mathbb{P}(TH) + \mathbb{P}(HT)} = \frac{p(1-p)}{(1-p)p + p(1-p)} \\ = \quad & \frac{p(1-p)}{2p(1-p)} = \frac{1}{2} = 50\%\end{aligned}$$

Therefore, we have shown that

$$\mathbb{P}(result = heads) = \frac{1}{2} = 50\% = \mathbb{P}(result = tails)$$

so the result is equally likely to be either heads or tails.

(b) **No.**

Under this simpler procedure, the result is heads $\iff$ the last coin flip is heads and all of the previous coin flips are tails. Therefore,

$$\mathbb{P}(result = heads) = \mathbb{P}(\{T^i H : i \geq 1\} = \mathbb{P}(TH \cup T^2 H \cup ...)$$

All of these outcomes $T^i H$ are mutually disjoint, so

$$\mathbb{P}(result = heads) = \sum_{i=1}^{\infty} \mathbb{P}(T^i H)$$

Since the coin flips are all mutually independent, we know $\mathbb{P}(T^i H) = \mathbb{P}(T)^i \mathbb{P}(H) = (1-p)^i p$, so

$$\mathbb{P}(result = heads) = \sum_{i=1}^{\infty} (1-p)^i p = p(1-p) \sum_{i=0}^{\infty} (1-p)^i = p(1-p) \frac{1}{1-(1-p)} = \frac{p(1-p)}{p} = 1 - p$$

We can note that this result is the same as $\mathbb{P}(T)$, since as soon as we flip tails on our first flip, the result in any finite game will be heads.

Similarly, the result under this simpler procedure is tails $\iff$ the last coin flip is tails and all of the previous coin flips are heads. Therefore,

$$\mathbb{P}(result = tails) = \mathbb{P}(\{H^i T : i \geq 1\} = \mathbb{P}(HT \cup H^2 T \cup ...)$$

All of these outcomes $H^i T$ are mutually disjoint, so

$$\mathbb{P}(result = tails) = \sum_{i=1}^{\infty} \mathbb{P}(H^i T)$$

Since the coin flips are all mutually independent, we know $\mathbb{P}(H^i T) = \mathbb{P}(H)^i \mathbb{P}(T) = p^i (1-p)$, so

$$\mathbb{P}(result = tails) = \sum_{i=1}^{\infty} p^i (1-p) = p(1-p) \sum_{i=0}^{\infty} p^i = p(1-p) \frac{1}{1-p} = p$$

This result similarly is the same as $\mathbb{P}(H)$, since as soon as we flip heads on the first flip, the result in any finite game must be tails.

Therefore, if we have a biased coin with $p \neq \frac{1}{2}$, then, under this simpler procedure,

$$\mathbb{P}(result = heads) = 1 - p \neq p = \mathbb{P}(result = tails)$$

so the results of heads and tails are not equally likely with a biased coin under the simpler procedure.

2. (Ross P3.76) Suppose that each child born to a couple is equally likely to be a boy or a girl, independently of the sex distribution of the other children in the family. For a couple having 5 children, compute the probabilities of the following events:

   (a) All children are of the same sex.

   (b) The 3 oldest are boys and the others girls.

   (c) Exactly 3 are boys.

   (d) The 2 oldest are girls.

   (e) There is at least 1 girl.

   Solution.
   Since there are two choices (boy or girl) for the gender of each of the 5 children, and these choices are all independent of each other, the sample space, S, has a size of

$$|S| = 2^5 = 32$$

   Since each child is equally likely to be a boy or a girl, each of these 32 outcomes is equally likely.

   (a) For all children to be of the same sex, they must either all be girls or all be boys. There is exactly one outcome in which all 5 children are girls, which is (G, G, G, G, G). Similarly, there is exactly one outcome in which all 5 children are boys, which is (B, B, B, B, B).
   Let A = the event that all 5 children are of the same sex. Then $|A| = 2$. Since all outcomes are equally likely, the probability that all 5 children are of the same sex is

$$\mathbb{P}(A) = \frac{|A|}{|S|} = \frac{2}{32} = \frac{1}{16} = 6.25\%$$

   (b) For the oldest 3 children to be boys and the others to be girls, we must have exactly 3 boys followed by exactly 2 girls. The only outcome in which this happens is (B, B, B, G, G), as changing the gender of any of the eldest three children will violate the condition that the eldest three are boys, and changing the gender of one of the youngest two children will violate the condition that the others are girls.
   Let O = the event that the oldest three are boys and the others girls. Then $|O| = 1$, so the probability that the oldest three are boys and the others are girls is

$$\mathbb{P}(O) = \frac{|O|}{|S|} = \frac{1}{32} = 3.125\%$$

   (c) For exactly 3 of the 5 children to be boys, we can select any 3 out of the 5 children to be boys, leaving the remaining 2 to be girls. This can be done in exactly $\binom{5}{3} = 10$ ways.
   Let T = the event that exactly 3 of the 5 children are boys. Then $|T| = 10$, so the probability that exactly 3 of the 5 children are boys is

$$\mathbb{P}(T) = \frac{|T|}{|S|} = \frac{10}{32} = \frac{5}{16} = 31.25\%$$

   (d) For the oldest 2 to be girls, the youngest 3 can have any sequence of genders. Since there are 2 independent, equally likely choices for the gender of each of these youngest 3, there are exactly $2^3 = 8$ equally likely ways for the oldest 2 to be girls. We could encode these as $(G, G, \{G, B\}^3)$.
   Let L = the event that the 2 oldest children are girls. Then $|L| = 8$, so the probability that the oldest 2 are girls is

$$\mathbb{P}(L) = \frac{|L|}{|S|} = \frac{8}{32} = \frac{1}{4} = 25\%$$

(e) Let F = the event that there is at least one girl. Then F^c = the event that there are exactly 0 girls among the 5 children. The only outcome in which this takes place is (B, B, B, B, B), so $|F^c| = 1$ and the probability that there are exactly 0 girls is

$$\mathbb{P}(F^c) = \frac{|F^c|}{|S|} = \frac{1}{32}$$

Since $\mathbb{P}(F) = 1 - \mathbb{P}(F^c)$, we know the probability that there is at least 1 girl is

$$\mathbb{P}(F) = 1 - \mathbb{P}(F^c) = 1 - \frac{1}{32} = \frac{31}{32} = 96.875\%$$

3. (Ross P3.89) Let $S = \{1, 2, \ldots, n\}$ and suppose that A and B are, independently, equally likely to be any of the 2^n subsets (including the null set and S itself) of S.

   (a) Show that
   $$P(A \subseteq B) = \left(\frac{3}{4}\right)^n.$$
   Hint: Let $N(B)$ denote the number of elements in B. Use
   $$P(A \subseteq B) = \sum_{i=0}^{n} P(A \subseteq B \mid N(B) = i) P(N(B) = i).$$

   (b) Show that $P(A \cap B = \varnothing) = \left(\frac{3}{4}\right)^n$.

Solution.

(a) Applying the hint, we find that
$$\mathbb{P}(A \subseteq B) = \sum_{i=0}^{n} \mathbb{P}(A \subseteq B \mid N(B) = i) \mathbb{P}(N(B) = i) \qquad (1)$$
Now, let's calculate $\mathbb{P}(A \subseteq B \mid N(B) = i)$ and $\mathbb{P}(N(B) = i)$ for all $0 \leq i \leq n$.

$\mathbb{P}(N(B) = i)$: Since B must be a subset of S, and there are 2^n subsets of S, there are 2^n equally likely possibilities for B. For $N(B) = i$, we need B to have exactly i elements from the n elements in S. There are exactly $\binom{n}{i}$ equally likely ways to choose which i of the n elements in S are also in B, so $|N(B) = i| = \binom{n}{i}$. Therefore, for all $0 \leq i \leq n$, the probability that B has exactly i elements is
$$\mathbb{P}(N(B) = i) = \frac{|N(B) = i|}{2^n} = \frac{\binom{n}{i}}{2^n}$$
$\mathbb{P}(A \subseteq B | N(B) = i)$: Since A and B are, independently, equally likely to be any of the 2^n subsets of S, there are 2^n equally likely possibilities for A. For A to be a subset of B, all elements in A must also be in B. Since we are given $N(B) = i$, we know that A cannot have more than i elements, so A must have j elements, where $0 \leq j \leq i$. Since there are i elements in B from which we choose elements in A, for each j, there are exactly $\binom{i}{j}$ ways to form A s.t. $A \subseteq B$ when $N(B) = i$. Therefore, $(A \subseteq B | N(B) = i)$ is a set with exactly
$$|A \subseteq B | N(B) = i| = \sum_{j=0}^{i} \binom{i}{j} = \sum_{j=0}^{i} \binom{i}{j} \cdot 1^j \cdot 1^{i-j} = (1+1)^i = 2^i$$
elements for all $0 \leq i \leq n$, with the second to last step following from the Binomial Theorem. Therefore, the probability that A is a subset of B given that $N(B) = i$ is
$$\mathbb{P}(A \subseteq B | N(B) = i) = \frac{2^i}{2^n}$$
Now, we can plug $\mathbb{P}(N(B) = i)$ and $\mathbb{P}(A \subseteq B | N(B) = i)$ into (1) to find that the total probability that A is a subset of B is
$$\begin{aligned} \mathbb{P}(A \subseteq B) = & \sum_{i=0}^{n} \mathbb{P}(A \subseteq B \mid N(B) = i) \mathbb{P}(N(B) = i) = \sum_{i=0}^{n} \frac{2^i}{2^n} \frac{\binom{n}{i}}{2^n} \\ = & \frac{1}{(2^n)^2} \sum_{i=0}^{n} \binom{n}{i} \cdot 2^i \cdot 1^{n-i} = \frac{1}{2^{2n}} (2+1)^n = \frac{3^n}{4^n} = \left(\frac{3}{4}\right)^n \end{aligned}$$
with the third to last step following from the Binomial Theorem. This completes the proof that
$$\mathbb{P}(A \subseteq B) = \left(\frac{3}{4}\right)^n$$

(b) Since A and B are selected independently and are both equally likely to be any of the 2^n subsets of S, we know there are 2^n equally likely possibilities for A for each of 2^n equally likely possibilities for B. This leaves $2^n \cdot 2^n = (2^n)^2 = 2^{2n} = 4^n$ equally likely possibilities for the combination of A and B.

Without loss of generality, suppose A has exactly i elements, where $0 \leq i \leq n$. There are $\binom{n}{i}$ equally likely ways to choose the i elements that form A from the n elements in S. In order for $A \cap B = \emptyset$, we can only form B by selecting from the remaining $n - i$ elements in $S - A$. Suppose B consists of exactly j of these $n - i$ remaining items, where $0 \leq j \leq n - i$. Then there are $\binom{n-i}{j}$ equally likely ways to choose which j elements from $S - A$ are in B. Summing over all possible values of j, we find there are

$$\sum_{j=0}^{n-i} \binom{n-i}{j} = \sum_{j=0}^{n-i} \binom{n-i}{j} \cdot 1^j \cdot 1^{n-i-j} = (1+1)^{n-i} = 2^{n-i}$$

total ways to choose B such that $A \cap B = \emptyset$, given that $N(A) = i$, with the second to last step following from the Binomial Theorem. Summing over all possible values of i, we find there are

$$\sum_{i=0}^{n} \binom{n}{i} 2^{n-i} = \sum_{i=0}^{n} \binom{n}{i} \cdot 2^{n-i} \cdot 1^i = (2+1)^n = 3^n$$

total equally likely ways to choose A and B such that $A \cap B = \emptyset$, with the second to last step following from the Binomial Theorem.
Since there are 4^n equally likely ways to choose A and B, the probability that $A \cap B = \emptyset$ is

$$\mathbb{P}(A \cap B = \emptyset) = \frac{3^n}{4^n} = (\frac{3}{4})^n$$

as required.

4. (Ross P4.7-4.8) Suppose that a die is rolled twice.

   (a) What are the possible values that the following random variables can take on:
      (1) the maximum value to appear in the two rolls;
      (2) the minimum value to appear in the two rolls;
      (3) the sum of the two rolls;
      (4) the value of the first roll minus the value of the second roll?

   (b) If the dice are fair, calculate the probabilities associated with the first and last random variables above.

   (c) Calculate the expected values of the first and last random variables above.

   (d) Calculate the variances of the first and last random variables above.

Solution.
For all parts of this problem, we have the sample space $S = \{1, 2, 3, 4, 5, 6\}^2$, so $|S| = 6^2 = 36$.
Also for all parts of this problem, if $a \in \{1, 2, 3, 4, 5, 6\}$ is rolled first and $b \in \{1, 2, 3, 4, 5, 6\}$ is rolled second, then we denote the result of the two rolls ab.

(a) (1) Let $X =$ the maximum value to appear in the two rolls.
Then $X : S \to \mathbb{R}$ where, if the result of the two rolls is ab, then

$$X(ab) = max(a, b)$$

We know that, for any outcome $ab \in S$, $X(ab) \in \{1, 2, 3, 4, 5, 6\}$ because the maximum value of the two rolls must be the value of one of the rolls, which must be some element in $\{1, 2, 3, 4, 5, 6\}$. We can quickly see that X can in fact take on all values of $\{1, 2, 3, 4, 5, 6\}$:

$$X(11) = 1 \qquad X(12) = 2 \qquad X(13) = 3 \qquad X(14) = 4 \qquad X(15) = 5 \qquad X(16) = 6$$

Therefore, the possible values of the maximum value to appear in the two rolls are all elements in $\{1, 2, 3, 4, 5, 6\}$

(2) Let $Y =$ the minimum value to appear in the two rolls.
Then $Y : S \to \mathbb{R}$ where, if the result of the two rolls is ab, then

$$Y(ab) = min(a, b)$$

We know that, for any outcome $ab \in S$, $Y(ab) \in \{1, 2, 3, 4, 5, 6\}$ because the minimum value of the two rolls must be the value of one of the rolls, which must be some element in $\{1, 2, 3, 4, 5, 6\}$. We can quickly see that Y can in fact take on all values of $\{1, 2, 3, 4, 5, 6\}$:

$$Y(16) = 1 \qquad Y(26) = 2 \qquad Y(36) = 3 \qquad Y(46) = 4 \qquad Y(56) = 5 \qquad Y(66) = 6$$

Therefore, the possible values of the minimum value to appear in the two rolls are all elements in $\{1, 2, 3, 4, 5, 6\}$

(3) Let $C =$ the sum of the two rolls.
Then $C : S \to \mathbb{R}$ where, if the result of the two rolls is ab, $C(ab) = a + b$.
The minimum value of C occurs when both dice rolls are minimal, so C is always at least $1 + 1 = 2$. The maximum value of C occurs when both dice rolls are maximal, so C is always at most $6 + 6 = 12$.
We can easily show that C can take on all integer values in $\{2, 3, ..., 11, 12\}$. Consider the result 11. $C(11) = 2$. Until the second coin has value 6, add one to the second coin. This will increase the value of C by 1 each time, and shows the possibility of all integer values of C from $2 \to 7$. Once the second coin has value 6, add one to the first coin until it has value 6. This

will continue to increase C by 1 each time until $C = 12$, showing that C can indeed take all values from $\{2, 3, ..., 11, 12\}$:

$C(11) = 2$ $C(12) = 3$ $C(13) = 4$ $C(14) = 5$ $C(15) = 6$ $C(16) = 7$
$C(26) = 8$ $C(36) = 9$ $C(46) = 10$ $C(56) = 11$ $C(66) = 12$

Therefore, the possible values of the sum of the two rolls are all elements in $\{2, 3, ..., 11, 12\}$

(4) Let M = the value of the first roll minus the value of the second roll.
Then $M : S \to \mathbb{R}$ where, if the result of the two rolls is ab, then $M(ab) = a - b$.
The minimum value of M occurs when a is minimal and b is maximal, so M is always at least $1 - 6 = -5$. The maximum value of M occurs when a is maximal and b is minimal, so M is always at most $6 - 1 = 5$.
We can easily show that M can take on all integer values in $\{-5, -4, ..., 4, 5\}$. Consider the minimal result 16. Until the second coin has value 1, subtract one from the second coin. This will increase the value of M by one each time, and shows the possibility of all integer values of M from $-5 \to 0$. Once the second coin has value 1, add 1 to the first coin until it has value 6. This will continue to increase M by 1 each time until $M = 6 - 1 = 5$, showing that M can indeed take all values from $\{-5, -4, ..., 4, 5\}$:

$M(16) = -5$ $M(15) = -4$ $M(14) = -3$ $M(13) = -2$ $M(12) = -1$ $M(11) = 0$
$M(21) = 1$ $M(31) = 2$ $M(41) = 3$ $M(51) = 4$ $M(61) = 5$

Therefore, the possible values of the first roll minus the second roll are all elements in $\{-5, -4, ..., 4, 5\}$

(b) We need to calculate $\mathbb{P}(X = x)$ for all $x \in \{1, 2, 3, 4, 5, 6\}$ and $\mathbb{P}(M = m)$ for all $m \in \{-5, -4, ..., 4, 5\}$.

$\mathbb{P}(X)$:

(i) If $X = 1$, then both rolls must be 1, so $(X = 1) = \{11\}$. Since all outcomes are equally likely, the probability that the maximum value to appear in the two rolls is 1 is

$$\mathbb{P}(X = 1) = \frac{|X = 1|}{|S|} = \frac{1}{36} \approx 2.78\%$$

(ii) If $X = 2$, then one roll must be a 2, and the other can be no larger than 2, so $(X = 2) = \{12, 21, 22\}$. Since all outcomes are equally likely, the probability that the maximum value to appear in the two rolls is 2 is

$$\mathbb{P}(X = 2) = \frac{|X = 2|}{|S|} = \frac{3}{36} = \frac{1}{12} \approx 8.33\%$$

(iii) If $X = 3$, then one roll must be a 3, and the other can be no larger than 3, so $(X = 3) = \{13, 31, 23, 32, 33\}$. Since all outcomes are equally likely, the probability that the maximum value to appear in the two rolls is 3 is

$$\mathbb{P}(X = 3) = \frac{|X = 3|}{|S|} = \frac{5}{36} \approx 13.89\%$$

(iv) If $X = 4$, then one roll must be a 4, and the other can be no larger than 4, so $(X = 4) = \{14, 41, 24, 42, 34, 43, 44\}$. Since all outcomes are equally likely, the probability that the maximum value to appear in the two rolls is 4 is

$$\mathbb{P}(X = 4) = \frac{|X = 4|}{|S|} = \frac{7}{36} \approx 19.44\%$$

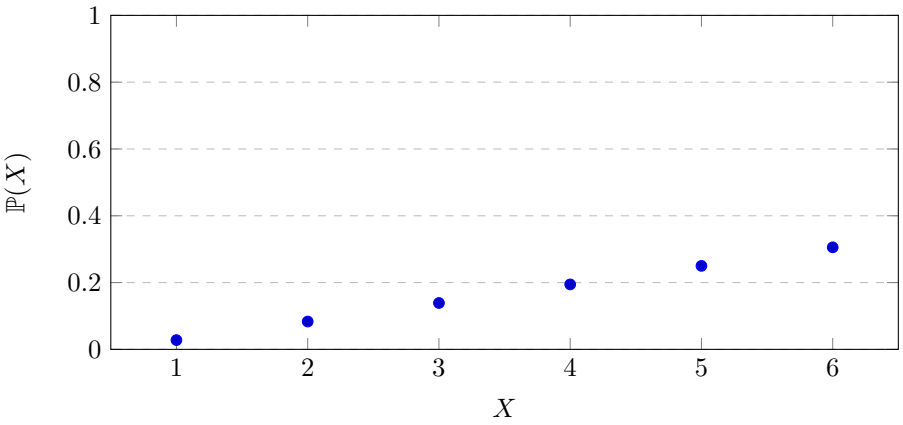
(v) If $X = 5$, then one roll must be a 5, and the other can be no larger than 5, so $(X = 5) = \{15, 51, 25, 52, 35, 53, 45, 54, 55\}$. Since all outcomes are equally likely, the probability that the maximum value to appear in the two rolls is 5 is

$$\mathbb{P}(X = 5) = \frac{|X = 5|}{|S|} = \frac{9}{36} = \frac{1}{4} = 25\%$$

(vi) If $X = 6$, then one roll must be a 6, and the other can be no larger than 6, so $(X = 6) = \{16, 61, 26, 62, 36, 63, 46, 64, 56, 65, 66\}$. Since all outcomes are equally likely, the probability that the maximum value to appear in the two rolls is 6 is

$$\mathbb{P}(X = 6) = \frac{|X = 6|}{|S|} = \frac{11}{36} \approx 30.56\%$$

We can plot these results to show the probability distribution of X.



$\mathbb{P}(M)$:

(i) If $M = -5$, then the first roll must be a 1 and the second roll must be a 6, so $(M = -5) = \{16\}$. Since all outcomes are equally likely, the probability that the value of the first roll minus the value of the second roll is -5 is

$$\mathbb{P}(M = -5) = \frac{|M = -5|}{|S|} = \frac{1}{36} \approx 2.78\%$$

(ii) If $M = -4$, then the value of the first roll minus the value of the second roll has to be -4, so $(M = -4) = \{15, 26\}$. Since all outcomes are equally likely, the probability that the value of the first roll minus the value of the second roll is -4 is

$$\mathbb{P}(M = -4) = \frac{|M = -4|}{|S|} = \frac{2}{36} = \frac{1}{18} \approx 5.56\%$$

(iii) If $M = -3$, then the value of the first roll minus the value of the second roll has to be -3, so $(M = -3) = \{14, 25, 36\}$. Since all outcomes are equally likely, the probability that the value of the first roll minus the value of the second roll is -3 is

$$\mathbb{P}(M = -3) = \frac{|M = -3|}{|S|} = \frac{3}{36} = \frac{1}{12} \approx 8.33\%$$

(iv) If $M = -2$, then the value of the first roll minus the value of the second roll has to be -2, so $(M = -2) = \{13, 24, 35, 46\}$. Since all outcomes are equally likely, the probability that the value of the first roll minus the value of the second roll is -2 is

$$\mathbb{P}(M = -2) = \frac{|M = -2|}{|S|} = \frac{4}{36} = \frac{1}{9} \approx 11.11\%$$

(v) If $M = -1$, then the value of the first roll minus the value of the second roll has to be -1, so $(M = -1) = \{12, 23, 34, 45, 56\}$. Since all outcomes are equally likely, the probability that the value of the first roll minus the value of the second roll is -1 is

$$\mathbb{P}(M = -1) = \frac{|M = -1|}{|S|} = \frac{5}{36} \approx 13.89\%$$

(vi) If $M = 0$, then the value of the first roll must equal the value of the second roll, so $(M = 0) = \{11, 22, 33, 44, 55, 66\}$ Since all outcomes are equally likely, the probability that the value of the first roll minus the value of the second roll is 0 is

$$\mathbb{P}(M = 0) = \frac{|M = 0|}{|S|} = \frac{6}{36} = \frac{1}{6} \approx 16.67\%$$

(vii) If $M = 1$, we can simply flip the order of all outcomes from $M = -1$, and then $a - b = -1$ becomes $b - a = 1$, so $(M = 1) = \{21, 32, 43, 54, 65\}$. Since all outcomes are equally likely, the probability that the value of the first roll minus the value of the second roll is 1 is

$$\mathbb{P}(M = 1) = \frac{|M = 1|}{|S|} = \frac{5}{36} \approx 13.89\%$$

(viii) If $M = 2$, we can simply flip the order of all outcomes from $M = -2$, and then $a - b = -2$ becomes $b - a = 2$, so $(M = 2) = \{31, 42, 53, 64\}$. Since all outcomes are equally likely, the probability that the value of the first roll minus the value of the second roll is 2 is

$$\mathbb{P}(M = 2) = \frac{|M = 2|}{|S|} = \frac{4}{36} = \frac{1}{9} \approx 11.11\%$$

(ix) If $M = 3$, we can simply flip the order of all outcomes from $M = -3$, and then $a - b = -3$ becomes $b - a = 3$, so $(M = 3) = \{41, 52, 63\}$. Since all outcomes are equally likely, the probability that the value of the first roll minus the value of the second roll is 3 is

$$\mathbb{P}(M = 3) = \frac{|M = 3|}{|S|} = \frac{3}{36} = \frac{1}{12} \approx 8.33\%$$

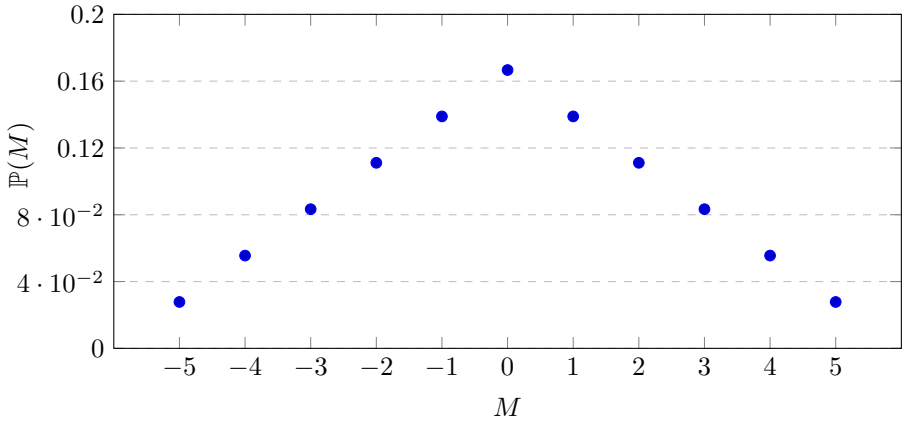
(x) If $M = 4$, we can simply flip the order of all outcomes from $M = -4$, and then $a - b = -4$ becomes $b - a = 4$, so $(M = 4) = \{51, 62\}$. Since all outcomes are equally likely, the probability that the value of the first roll minus the value of the second roll is 4 is

$$\mathbb{P}(M = 4) = \frac{|M = 4|}{|S|} = \frac{2}{36} = \frac{1}{18} \approx 5.56\%$$

(xi) If $M = 5$, then the first flip must be 6, and the second flip must be 1, so $(M = 5) = \{61\}$. Since all outcomes are equally likely, the probability that the value of the first roll minus the value of the second roll is 5 is

$$\mathbb{P}(M = 5) = \frac{|M = 5|}{|S|} = \frac{1}{36} \approx 2.78\%$$

We can plot these results to show the probability distribution of M:



(c) We will use the fact that, for any event X with values in I, we have

$$E[X] = \sum_{k \in I} k\mathbb{P}(X = k) \qquad (1)$$

Applying (1) to X, we find that the expected value for the maximum value to appear in the two dice rolls is

$$\begin{aligned} E[X] = & \sum_{k=1}^{6} k\mathbb{P}(X = k) = 1 \cdot \frac{1}{36} + 2 \cdot \frac{3}{36} + 3 \cdot \frac{5}{36} + 4 \cdot \frac{7}{36} + 5 \cdot \frac{9}{36} + 6 \cdot \frac{11}{36} \\ = & \frac{1 + 6 + 15 + 28 + 45 + 66}{36} = \frac{161}{36} \approx 4.47 \end{aligned}$$

Applying (1) to M, we find that the expected value for the value of the first roll minus the value of the second roll is

$$\begin{aligned} E[M] = & \sum_{k=-5}^{5} k\mathbb{P}(M = k) \\ = & -5 \cdot \frac{1}{36} - 4 \cdot \frac{2}{36} - 3 \cdot \frac{3}{36} - 2 \cdot \frac{4}{36} - 1 \cdot \frac{5}{36} + 0 \cdot \frac{6}{36} \\ & + 1 \cdot \frac{5}{36} + 2 \cdot \frac{4}{36} + 3 \cdot \frac{3}{36} + 4 \cdot \frac{2}{36} + 5 \cdot \frac{1}{36} \\ = & \frac{-5 - 8 - 9 - 8 - 5 + 0 + 5 + 8 + 9 + 8 + 5}{36} = \frac{0}{36} = 0 \end{aligned}$$

(d) We will use the fact that, for any event X with values in I,

$$Var(X) = E[X^2] - (E[X])^2 = \sum_{k \in I} k^2\mathbb{P}(X = k) - (E[X])^2 \qquad (2)$$

$Var(X)$: We can directly compute that

$$\begin{aligned} \sum_{k \in I} k^2\mathbb{P}(X = k) = & \sum_{k=1}^{6} k^2\mathbb{P}(X = k) \\ = & 1^2 \cdot \frac{1}{36} + 2^2 \cdot \frac{3}{36} + 3^2 \cdot \frac{5}{36} + 4^2 \cdot \frac{7}{36} + 5^2 \cdot \frac{9}{36} + 6^2 \cdot \frac{11}{36} \\ = & \frac{1 + 4 \cdot 3 + 9 \cdot 5 + 16 \cdot 7 + 25 \cdot 9 + 36 \cdot 11}{36} \\ = & \frac{1 + 12 + 45 + 112 + 225 + 396}{36} = \frac{791}{36} \end{aligned}$$

Combining this result with our result for $E[X]$ from part (c) and plugging into (2), we find that the variance of X is

$$Var(X) = \frac{791}{36} - (\frac{161}{36})^2 = \frac{28476 - 25921}{1296} = \frac{2555}{1296} \approx 1.97$$

$Var(M)$: We can directly compute that

$$\begin{aligned} \sum_{k \in I} k^2\mathbb{P}(M = k) = & \sum_{k=-5}^{5} k^2\mathbb{P}(M = k) \\ = & (-5)^2 \cdot \frac{1}{36} + (-4)^2 \cdot \frac{2}{36} + (-3)^2 \cdot \frac{3}{36} + (-2)^2 \cdot \frac{4}{36} + (-1)^2 \cdot \frac{5}{36} + 0^2 \cdot \frac{6}{36} \\ & + 1^2 \cdot \frac{5}{36} + 2^2 \cdot \frac{4}{36} + 3^2 \cdot \frac{3}{36} + 4^2 \cdot \frac{2}{36} + 5^2 \cdot \frac{1}{36} \\ = & \frac{25 + 32 + 27 + 16 + 5 + 0 + 5 + 16 + 27 + 32 + 25}{36} = \frac{210}{36} = \frac{35}{6} \approx 5.83 \end{aligned}$$

Combining this result with our result for $E[M]$ from part (c) and plugging into (2), we find that the variance of M is

$$Var(M) = \frac{35}{6} - 0^2 = \frac{35}{6} \approx 5.83$$

5. The October 1st, 2022 drawing of the Philippine Grand Lotto jackpot had an astonishing 433 winners. The jackpot was worth roughly $4 million USD.

   (a) To play the game, "LOTTO 6-55," you buy a ticket and pick 6 distinct numbers from $1, 2, \ldots, 55$. The lottery officials randomly do the same on the night of the drawing and come up with their own set of 6 distinct numbers from $1, 2, \ldots, 55$. You win the jackpot if your set of 6 numbers matches the lottery officials' set. Compute the probability that you win the jackpot if you buy a single ticket.

   (b) Suppose n people played. What is the probability of getting exactly 433 winners? Be sure to state any assumptions that go into your calculation.

   (c) Suppose that $n = 10{,}000{,}000$ people played. Use the Poisson approximation to compute the probability that there were 433 winners.

   (d) Dr. Guido David of the University of the Philippines posted an analysis of the situation online. See Figure 1. Do you agree with the analysis presented? What probabilistic notions are behind Dr. David's calculations?

   (e) According to the New York Post, "Philippine Senate Minority Leader Aquilino 'Koko' Pimentel called for official hearings into the 'strange and unusual' result." Do you believe hearings are warranted given the available evidence?

   Solution.

   (a) Let $W =$ the event that you win the jackpot after buying a single ticket.
   There are exactly $\binom{55}{6}$ equally likely ways to randomly select 6 distinct numbers from $1, 2, ..., 55$. Only one of these $\binom{55}{6}$ distinct combinations of 6 numbers is identical to the combination drawn by the lottery officials. Therefore, since all outcomes are equally likely, the probability that you win the jackpot if you buy a single ticket is

$$\mathbb{P}(E) = \frac{1}{\binom{55}{6}} \approx 0.00000345\%$$

   (b) Let $X =$ the number of people that win the jackpot when n people each buy one ticket. We want to find $\mathbb{P}(X = 433)$
   We assume that the choices of 6 distinct numbers are independent for each of the n people playing the game, and that each of the n people randomly choose their combinations of 6 numbers. Then we can consider n people playing the game as n independent Bernoulli trials, where $p = \mathbb{P}(\text{success}) = \frac{1}{\binom{55}{6}}$ and $q = \mathbb{P}(\text{failure}) = 1 - p$. Then $X \sim Binomial(n, p)$, so we know that the probability that exactly 433 out of the n people win the jackpot is

$$\mathbb{P}(X = 433) = p(433) = \binom{n}{433}\Big(\frac{1}{\binom{55}{6}}\Big)^{433}\Big(1 - \frac{1}{\binom{55}{6}}\Big)^{n-433}$$

   (c) Since $X \sim Binomial(n, p)$, $n = 10,000,000$ is very large, and $p = \frac{1}{\binom{55}{6}}$ is very small, the Poisson approximation guarantees that, with $\lambda = np = \frac{10{,}000{,}000}{\binom{55}{6}}$,

$$\mathbb{P}(X = 433) \approx e^{-\lambda}\frac{\lambda^{433}}{433!} = e^{-\frac{10{,}000{,}000}{\binom{55}{6}}}\frac{\Big(\frac{10{,}000{,}000}{\binom{55}{6}}\Big)^{433}}{433!} \approx 0 = 0\%$$

   Thus, using Poisson approximation, the probability that there were exactly 433 winners if $n = 10,000,000$ people played the game is approximately 0%.

(d) I agree with Dr. David's analysis.
Dr. David's result for the probability that you win the jackpot if you buy one ticket matches my result from part (a). This result relies on the assumption that you randomly select the numbers on your ticket.
I also agree with Dr. David's calculation of the probability that you do not win the jackpot with your ticket, which relies on the probabilistic notion of complements, specifically that $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$. Dr. David's calculation of the likelihood that at least one person wins the jackpot out of N people also relies on complements, as well as the assumption that all N people choose their numbers randomly and independently of one another.
Dr. David's line of reasoning for why randomly choosing $9, 18, 27, 36, 45, 54$ is just as likely/unlikely as choosing $3, 14, 15, 9, 26, 54$ is sound, but only under the assumption that the numbers are randomly chosen. Dr. David addresses this in the last paragraph of the analysis, recognizing how human behavior could make people more likely to choose a nice pattern like $9, 18, 27, 36, 45, 54$ than a random combination like $14, 3, 8, 43, 44, 36$.
Dr. David's calculation of the extremely low likelihood of exactly 433 jackpot winners out of 10 million randomly chosen tickets also matches my result from part (c). Dr. David sets the calculation up like my result from part (b), using the binomial theorem and representing the number of jackpot winners out of n tickets as a binomial random variable. Similar to how we use Poissant approximation in part (c), Dr. David uses an approximation on the size of $\binom{10,000,000}{433}$ to estimate to calculate the probability of exactly 433 jackpot winners out of 10,000,000 randomly chosen tickets.

(e) Given the available evidence, I believe the hearings are warranted. Since the upper limit on the estimates for the number of tickets was 10,000,000, the probability that exactly 433 people win the jackpot out of 10,000,000 tickets should be an upper bound on the probability of the event that actually occurred, assuming that the tickets are selected randomly. Therefore, since my result from part (c) shows that this probability is approximately 0%, the result of the October 1st drawing is indeed quite 'strange and unusual.' However, as Dr. David addresses and we address in part (d), the assumption that all tickets are selected randomly underlies this probability calculation. Since humans tend to prefer patterns to randomness, it is not entirely safe to assume that the 10,000,000 tickets were chosen randomly. Since the winning combination happened to be such a nice combination, $9, 18, 27, 36, 45, 54$, it is quite possible that significantly more than $\frac{1}{\binom{55}{6}}$ of the lottery tickets chose this combination without any corruption or cheating. However, since the specific ticket-choosing behavior of the population cannot be determined, the only way to accurately calculate the probability of a certain number of jackpot winners is to assume randomly selected tickets. Therefore, since the probability calculation of the observed outcome, which is valid under its assumptions, is so incredibly low, the hearings are warranted. However, since one of these underlying assumptions is not entirely safe, more evidence must be collected to determine with certainty whether the observed outcome of 433 winners involved cheating or happened by chance legitimately. It is also important to note that the calculated probability of 433 winners does not consider the possibility of more than 433 winners, so the probability of an outcome at least as extreme as the one observed on October 1st is slightly higher than the probability calculated in part (c). However, this cumulative probability is still ≈ 0, so the hearings are still warranted to collect further evidence.

6-55 LOTTO PROBABILITIES

by: Dr. Guido David

The estimated odds of 433 6-55 winners

When you play the LOTTO 6-55, you have a probability of winning of 1 in 28989675. There are 55 numbers (from 1 to 55) and the total number of combinations of 6 unique numbers out of 5 is ${}_{55}C_6 = 55! \div (6! \times 49!) = 28989675$. So that means you have roughly a 1 in 29 million chance of hitting the lottery. Let us denote this number by p, i.e. p = 1/28989675 = 0.000000034495, or p = 3.45e-8 in scientific notation. This means 3.45 × 0.00000001 or 3.45 × 1/100000000, i.e. 1 divided by a number with 8 zeros. Now we let q = 1 - p = 0.99999996550496. This is your odds of not winning the lotto.

How often is the lotto won? It depends on how many bettors there are. Assuming there are N = 1,000,000 bettors, the probability of at least one lotto winner is 1 minus the probability of no winners. This number is $1 - q^N = 0.0339$, i.e. roughly 3% of the time there will be at least one winner. If the number of bettors increase to N = 10,000,000, the probability of at least one lotto winner increases to 29%, nearly ten times compared to just 1,000,000 bettors. If the number of bettors increase to N = 100,000,000, nearly the population of the country, the probability of at least one lotto winner increases to 97%, i.e. at least one winner is almost guaranteed with this many bettors.

What are the chances that the outcome is specifically the numbers: 9, 18, 27, 36, 45, 54, which form a mathematical sequence (all multiples of 9)? The answer is that it is the same as any other particular outcome, for example 3 14 15 9 26 54 (for those curious, I just took those from the digits of the mathematical constant pi). The probability of this occurring is still p.

The next question is what are the chances of 433 lotto winners? We use the binomial theorem. We note that $p^{433} = 4.88e\text{-}33$ accounts for the winners. While we don't know the number of bets placed, historical data on the chances of winning would put this as somewhere between 1 million and 10 million. Assuming 10 million bets, then $q^{10000000-433} = 0.71$. This would gives us $p^{433}\ q^{10000000-433} = 3.46e\text{-}3300$. However, we have to multiply this by the combinations of people winning, given by ${}_{10000000}C_{433}$. This number is too large to be calculated even by computers, but we can estimate the numerator to be (1e7)^(433) = 1e3031. The denominator is 433! = 1.85e955, so that ${}_{10000000}C_{433}$ is approximately 5.4e2075. This gives us the probability of 433 winners out of 10 million bets to be 1.87e-1224. In words, this is once out of a number with 1224 zeros. To compare, a googol has 100 zeros. The number of molecules in the known universe has 80 zeros. The age of the known universe, in seconds, is about 4.32e17 (17 zeros).

There are some circumstances that may increase the probability of this happening. The fact that the winning combination is a a nice mathematical combination consisting of multiples of 9, compared to a combination with no pattern, may make it more likely someone would bet on such a combination.

Figure 1: Dr. Guido David's analysis of the 433 LOTTO 6-55 winners.

6. (a) Suppose X is a discrete random variable with probability generating function $G_X(t)$. Show that

$$G_X(t) = E[t^X].$$

(b) Suppose A is a randomly chosen subset of $[n]$, where each subset is equally likely. Let $X = \#A$ and let Y be the number of subsets of A. Compute the expected value of Y.

Solution.

(a) By the definition of the probability generating function, we know that if I is the set of values for which the discrete random variable X has nonzero probabilities, then

$$G_X(t) = \sum_{k \in I} p(k) t^k \qquad (1)$$

where $p(k)$ is a valid probability mass function.

Now, let's think about the right hand side of the identity we want to prove. By the definition of expected value, we know that

$$E[X] = \sum_{k \in I} k p(k)$$

For each possible value of X, we multiply that value by its corresponding probability and add up the results to find the expected value of X as a whole.
We want to compute $E[t^X]$. Now, for each $k \in I$, the value of t^X is t^k instead of k. However, since t is constant with respect to X, $\mathbb{P}(t^X = t^k)$ still equals $p(k)$, so the probability mass function of t^X is the same as that of X itself. Therefore, by multiplying each possible value of t^X by its corresponding probability and adding up the results, just like we did to find $E[X]$, we find

$$E[t^X] = \sum_{k \in I} p(k) t^k \qquad (2)$$

Comparing (1) and (2), we find they are equivalent, which completes the proof that

$$G_X(t) = E[t^X]$$

We could also note that

$$\begin{aligned} E[t^X] = & \sum_i \mathbb{P}(t^X = i) = \sum_i \sum_{k \text{ s.t. } g(k)=i} \mathbb{P}(X = k) = \sum_{k \in I} \sum_{i = t^k} i \mathbb{P}(X = k) \\ = & \sum_{k \in I} t^k \mathbb{P}(X = k) = \sum_{k \in I} p(k) t^k = G_X(t) \end{aligned}$$

for a more direct, computational proof.

(b) First, we need to determine all values for which Y has nonzero probabilities. For any X, there are 2^X subsets of A. This follows from the fact that there are $\binom{X}{i}$ subsets of size X for all $0 \le i \le X$, for a total of

$$\sum_{i=0}^{X} \binom{X}{i} = \sum_{i=0}^{X} \binom{X}{i} \cdot 1^i \cdot 1^{X-i} = (1+1)^X = 2^X$$

subsets, with the second to last step following from the Binomial Theorem.
Since A is a randomly chosen subset of $[n]$, we know $0 \le X = |A| \le n$. Therefore, we know the values for which Y has nonzero probabilities are $\{2^i | 0 \le i \le n\}$.
By the definition of expected value, we know that

$$E[Y] = \sum_{k=0}^{n} 2^k \mathbb{P}(X = k) \qquad (3)$$

There are exactly $\binom{n}{k}$ equally likely ways to choose a subset of size k from [n]. Since there are 2^n total subsets of [n], the probability that A, a randomly selected one, has exactly k elements is

$$\mathbb{P}(X = k) = \frac{\binom{n}{k}}{2^n}$$

for all $0 \leq k \leq n$. Plugging this into (3), we find the expected number of subsets of A is

$$\begin{aligned} E[Y] = & \sum_{k=0}^{n} 2^k \frac{\binom{n}{k}}{2^n} \\ = & \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k} \cdot 2^k \cdot 1^{n-k} \\ = & \frac{(2+1)^n}{2^n} = \left(\frac{3}{2}\right)^n \end{aligned}$$

with the second to last step following from the Binomial Theorem.
Thus, the expected value of Y is

$$E[Y] = \left(\frac{3}{2}\right)^n$$

7. Use calculations with probability generating functions to complete the following table.

Distribution	Parameter(s)	PMF	PGF	μ	σ^2
Binomial	n, p	$\binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$	$(pt + (1-p))^n$	np	$np(1-p)$
Geometric	p	$(1-p)^{k-1}p$	$\frac{pt}{1-(1-p)t}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative binomial	r, p	$\binom{k-1}{r-1} \cdot (1-p)^{k-r} \cdot p^r$	$(\frac{pt}{1-(1-p)t})^r$	$\frac{r}{p}$	$\frac{(1-p)r}{p^2}$
Poisson	λ	$e^{-\lambda}\frac{\lambda^k}{k!}$	$e^{(t-1)\lambda}$	λ	λ
Discrete uniform	a, b	$\frac{1}{b-a+1}$	$\frac{t^a}{b-a+1} \cdot \frac{1-t^{b-a+1}}{1-t}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$

Solution.

(i) Binomial: From lecture, we know the probability mass function of a random variable $X \sim Binomial(n, p)$ is

$$p(k) = \mathbb{P}(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

This directly implies that the probability generating function of $X \sim Binomial(n, p)$ is

$$G_X(t) = G(t) = \sum_{k=0}^{n} p(k)t^k = \sum_{k=0}^{n} \binom{n}{k} \cdot p^k t^k \cdot (1-p)^{n-k} = (pt + (1-p))^n$$

with the last step following from the Binomial Theorem.
Since $E[X] = G'(1)$, and the derivative of the PGF is

$$\frac{d}{dt}(pt + (1-p))^n = n(pt + (1-p))^{n-1}p$$

we can evaluate at $t = 1$ to find that the expected value of $X \sim Binomial(n, p)$ is

$$\mu_X = E[X] = n(p(1) + (1-p))^{n-1}p = np(1)^{n-1} = np$$

We know that $\sigma^2 = Var(X) = E[X^2] - E[X]^2$, and we already found $E[X] = np$, so we just need to compute $E[X^2]$. Note that

$$E[X^i] = \sum_{k=0}^{n} \binom{n}{k} k^i p^k (1-p)^{n-k} = \sum_{k=1}^{n} \binom{n}{k} k^i p^k (1-p)^{n-k}$$

Applying the fact that $k\binom{n}{k} = n\binom{n-1}{k-1}$, we find

$$\begin{aligned} E[X^i] = & \quad n\sum_{k=1}^{n}\binom{n-1}{k-1}k^{i-1}p^k(1-p)^{n-k} \\ = & \quad np\sum_{k=1}^{n}\binom{n-1}{k-1}k^{i-1}p^{k-1}(1-p)^{n-k} \\ = & \quad np\sum_{j=0}^{n-1}\binom{n-1}{j}(j+1)^{i-1}p^{j}(1-p)^{n-1-j} \\ = & \quad npE[(Y+1)^{i-1}] \end{aligned}$$

where $Y \sim Binomial(n-1,p)$. Therefore, setting $i=2$, we find

$$E[X^2] = npE[Y+1] = np((n-1)p+1)$$

since $E[Y+1] = E[Y]+1$. Plugging this into the equation for $\sigma^2 = Var(X)$, we find

$$\begin{aligned} \sigma^2 = Var(X) = & \quad E[X^2]-E[X]^2 = np((n-1)p+1) - n^2p^2 \\ = & \quad np(np-p+1) - n^2p^2 = n^2p^2 - np^2 + np - n^2p^2 \\ = & \quad np - np^2 = np(1-p) \end{aligned}$$

(ii) Geometric: From lecture, we know that the probability mass function of a random variable $T \sim Geometric(p)$ is

$$p(k) = \mathbb{P}(T=k) = (1-p)^{k-1}p$$

This directly implies that the probability generating function of $T \sim Geometric(p)$ is

$$\begin{aligned} G_T(t) = G(t) = & \quad \sum_{k=1}^{\infty}(1-p)^{k-1}t^kp = pt\sum_{k=0}^{\infty}(1-p)^kt^k \\ = & \quad pt\frac{1}{1-(1-p)t} = \frac{pt}{1-(1-p)t} = pt(1-(1-p)t)^{-1} \end{aligned}$$

We know that $E[T] = G'(1)$. We can take the derivative of $G_T(t)$ to find

$$\frac{d}{dt}\frac{pt}{1-(1-p)t} = p\cdot\frac{(1-(1-p)t)-(p-1)t}{(1-(1-p)t)^2} = \frac{p}{(1-(1-p)t)^t}$$

Plugging in $t=1$, we find that the expected value of $T \sim Geometric(p)$ is

$$E[T] = \mu = \frac{p}{(1-(1-p))^2} = \frac{p}{p^2} = \frac{1}{p}$$

We know that $\sigma^2 = Var(T) = E[T^2]-E[T]^2$, and we already found $E[T]$, so we just need to compute $E[T^2]$. Note that

$$\begin{aligned} E[T^2] = & \quad \sum_{k=1}^{\infty}k^2p(k) = \sum_{k=1}^{\infty}k^2(1-p)^{k-1}p \\ = & \quad \sum_{k=1}^{\infty}(k-1+1)^2(1-p)^{k-1}p \\ = & \quad \sum_{k=1}^{\infty}(k-1)^2(1-p)^{k-1}p + \sum_{k=1}^{\infty}2(k-1)(1-p)^{k-1}p + \sum_{k=1}^{\infty}(1-p)^{k-1}p \\ = & \quad \sum_{j=1}^{\infty}j^2(1-p)^jp + 2\sum_{j=1}^{\infty}j(1-p)^jp + 1 \\ = & \quad (1-p)E[T^2] + 2(1-p)E[T] + 1 \end{aligned}$$

This directly implies that

$$pE[T^2] = 2(1-p)E[T] + 1 = \frac{2(1-p)}{p} + 1$$

so

$$E[T^2] = \frac{2(1-p)+p}{p^2} = \frac{2-p}{p^2}$$

Plugging this into the equation for $\sigma^2 = Var(T)$, we find that the variance of $T \sim Geometric(p)$ is

$$\sigma^2 = Var(T) = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

(iii) Negative Binomial: From, lecture, we know the probability mass function of a $N \sim NegativeBinomial(r,p)$ is

$$p(k) = \mathbb{P}(N=k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r$$

We also found in lecture that the probability generating function of any $N \sim NegativeBinomial(r,p)$ is

$$\begin{aligned} G_N(t) = G(t) = & \quad \sum_{k\geq r}\binom{k-1}{r-1}(1-p)^{k-r}p^r t^k = (pt)^r(1-(1-p)t)-r \\ = & \quad \frac{(pt)^r}{(1-(1-p))^r} = (\frac{pt}{1-(1-p)t})^r \end{aligned}$$

To find $E[N]$, we will first find $E[N^k]$ using generating functions.

$$\begin{aligned} E[N^k] = & \quad \sum_{i=r}^{\infty}\binom{i-1}{r-1}i^k(1-p)^{i-r}p^r \\ = & \quad \frac{r}{p}\sum_{i=r}^{\infty}\binom{i}{r}i^{k-1}(1-p)^{n-r}p^{r+1} \text{ using } i\binom{i-1}{r-1} = r\binom{i}{r} \\ = & \quad \frac{r}{p}\sum_{m=r+1}^{\infty}\binom{m-1}{r}(m-1)^{k-1}(1-p)^{m-(r+1)}p^{r+1} \\ = & \quad \frac{r}{p}E[(M-1)^{k-1}] \end{aligned}$$

where $M \sim NegativeBinomial(1+r,p)$.
Plugging in $k=1$, we find that the expected value of $N \sim NegativeBinomial(r,p)$ is

$$E[N] = \frac{r}{p}E[(M-1)^0] = \frac{r}{p}$$

We could also note that, since $E[N] = G'(1)$, and the derivative of the PGF is

$$\frac{d}{dt}(\frac{pt}{1-(1-p)t})^r = \frac{r(\frac{pt}{1-(1-p)t})^r}{(p-1)t^2+t}$$

evaluating at $t=1$ yields

$$E[N] = G'(1) = \frac{r(\frac{p}{1-(1-p)})^r}{(p-1)+1} = \frac{r(\frac{p}{p})^r}{p} = \frac{r}{p}$$

as expected.
We know $\sigma^2 = Var(N) = E[N^2] - E[N]^2$, and we already calculated $E[N]$, so we just need to

compute $E[N^2]$.
Plugging $k = 2$ into the equation for $E[N^k]$, we find

$$E[N^2] = \frac{r}{p}E[(Y-1)] = \frac{r}{p}(\frac{r+1}{p} - 1) = \frac{r}{p}\frac{r+1-p}{p} = \frac{r^2+r-rp}{p^2}$$

Plugging this into the equation for $Var(N)$, we find that the variance of $N \sim NegativeBinomial(r,p)$ is

$$\sigma^2 = Var(N) = \frac{r^2+r-rp}{p^2} - \frac{r^2}{p^2} = \frac{r-rp}{p^2} = \frac{(1-p)r}{p^2}$$

(iv) Poisson: By the definition of a Poisson random variable, the probability mass function for any $P \sim Poisson(\lambda)$ is

$$p(k) = \mathbb{P}(P = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

for all $k \in \{0, 1, 2, ...\}$. This directly implies that the probability generating function of $P \sim Poisson(\lambda)$ is

$$G_P(t) = G(t) = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}t^k = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda\cdot t)^k}{k!}$$

Note: Using the Taylor Series expansion of e^x, we find

$$e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!}$$

So we know the probability generating function of $P \sim Poisson(\lambda)$ is

$$G(t) = e^{-\lambda}e^{\lambda\cdot t} = \frac{e^{\lambda\cdot t}}{e^{\lambda}} = e^{(t-1)\lambda}$$

By the definition of expected value, we know that

$$\mu = E(P) = \sum_{k=0}^{\infty} kp(k) = \sum_{k=0}^{\infty}\frac{ke^{-\lambda}\lambda^k}{k!}$$

Pulling out a factor of λ, cancelling the common factor of k, and discarding the $k = 0$ term which contributes 0 to the sum, we find that the expected value of $P \sim Poisson(\lambda)$ is

$$\mu = E(P) = \lambda\sum_{k=1}^{\infty}\frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{(j)!} = \lambda\cdot e^{-\lambda}\cdot e^{\lambda} = \lambda$$

with the second to last step following from the Taylor Series expansion of e^{λ}.
We could also note that, since $E[P] = G'(1)$ and the derivative of the PGF is

$$\frac{d}{dt}e^{(t-1)\lambda} = \lambda e^{(t-1)\lambda}$$

evaluating at $t = 1$ yields

$$E[P] = \lambda e^{1-1} = \lambda e^0 = \lambda$$

as expected.
We know that $\sigma^2 = Var(P) = E[P^2] - E[P]^2$, and we already calculated $E[P]$, so we just need to

calculate $E[P^2]$. Note that, by the definition of expected value

$$\begin{aligned} E[P^2] = & \sum_{k=0}^{\infty} \frac{k^2 e^{-\lambda}\lambda^k}{k!} = \lambda \sum_{k=1}^{\infty} \frac{k e^{-\lambda}\lambda^{k-1}}{(k-1)!} \\ = & \lambda \sum_{j=0}^{\infty} \frac{(j+1) e^{-\lambda}\lambda^{j}}{j!} = \lambda \left(\sum_{j=0}^{\infty} \frac{j e^{-\lambda}\lambda^{j}}{j!} + \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j}}{j!} \right) \\ = & \lambda(E[P] + G(1)) = \lambda(\lambda + 1) = \lambda^2 + \lambda \end{aligned}$$

Plugging this into the equation for $\sigma^2 = Var(P)$, we find that the variance of $P \sim Poission(\lambda)$ is

$$\sigma^2 = Var(P) = E[P^2] - E[P]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda = E[P]$$

Thus, both the variance and expected value of $P \sim Poisson(\lambda)$ equal λ.

(v) Discrete Uniform: Since all outcomes are equally likely, and there are $b - a + 1$ outcomes in $a, a+1, .., b$, we know that the probability mass function of any $D \sim DiscreteUniform(a, b)$ is

$$p(k) = \mathbb{P}(D = k) = \frac{1}{b - a + 1}$$

for all $k \in \{a, a+1, ..., b\}$ and $p(k) = 0$ for all other k.
This directly implies that the probability generating function of $D \sim DiscreteUniform(a, b)$ is

$$\begin{aligned} G_D(t) = G(t) = & \sum_{k=a}^{b} \frac{t^k}{b - a + 1} = \frac{1}{b - a + 1} \sum_{k=a}^{b} t^k = \frac{1}{b - a + 1} \sum_{k=0}^{b-a} t^{k+a} \\ = & \frac{t^a}{b - a + 1} \sum_{k=0}^{b-a} t^k = \frac{t^a}{b - a + 1} \cdot \frac{1 - t^{b-a+1}}{1 - t} \end{aligned}$$

By the definition of expected value, we know that

$$\mu = E[D] = \sum_{k=a}^{b} \frac{k}{b - a + 1}$$

Moving the $\frac{1}{b-a+1}$ outside the sum, we find that the expected value of $D \sim DiscreteUniform(a, b)$ is

$$\mu = E[D] = \frac{1}{b - a + 1} \sum_{k=a}^{b} k$$

For all $a, b \in \mathbb{Z}$,

$$\begin{aligned} \sum_{k=a}^{b} k = & \quad a + (a+1) + ... + (b-1) + b \\ = & \quad b + (b-1) + ... + (a+1) + a \end{aligned}$$

By adding vertically, each $a+i$ is paired with $b-i$ such that $a+i+b-i = a+b$. There are $b-a+1$ such pairs, so

$$2 \sum_{k=a}^{b} k = (b - a + 1)(a + b) \implies \sum_{k=a}^{b} k = \frac{(b - a + 1)(a + b)}{2}$$

Therefore, the expected value of $D \sim DiscreteUniform(a, b)$ is

$$\mu = E[D] = \frac{1}{b - a + 1} \sum_{k=a}^{b} k = \frac{1}{b - a + 1} \cdot \frac{(b - a + 1)(a + b)}{2} = \frac{a + b}{2}$$

We could also note that, since $E[D] = G'(1)$, and the derivative of the PGF is

$$\frac{d}{dt}\frac{t^a}{b-a+1}\cdot\frac{1-t^{b-a+1}}{1-t} = \frac{t^a(1-t^{b-a+1})}{(b-a+1)(1-t)^2} + \frac{at^{a-1}(1-t^{b-a+1})}{(b-a+1)(1-t)} - \frac{t^b}{1-t}$$

evaluating at $t = 1$ yields

$$\begin{aligned}
E[D] = \quad & G'(1) = \lim_{t\to 1}\frac{t^a(1-t^{b-a+1})}{(b-a+1)(1-t)^2} + \frac{at^{a-1}(1-t^{b-a+1})}{(b-a+1)(1-t)} - \frac{t^b}{1-t} \\
= \quad & \lim_{t\to 1}\frac{at^{a-1}(1-t^{b-a+1}) + t^a(-(b-a+1)t^{b-a})}{-2(b-a+1)(1-t)} \\
& + \frac{a(a-1)t^{a-2}(1-t^{b-a+1}) + at^{a-1}(-(b-a+1)t^{b-a})}{-(b-a+1)} - \frac{bt^{b-1}}{-1} \\
= \quad & \lim_{t\to 1}\frac{a(a-1)t^{a-2}(1-t^{b-a+1}) + at^{a-1}(-(b-a+1)t^{b-a}) - bt^{b-1}(b-a+1)}{2(b-a+1)} + a + b \\
= \quad & \frac{-a-b}{2} + a + b = \frac{-a-b+2a+2b}{2} = \frac{a+b}{2}
\end{aligned}$$

as expected.
We know that $\sigma^2 = Var(D) = E[D^2] - E[D]^2$.
Since $\sigma^2(D) = \sigma^2(D+x)$, where x is a constant, we can add $-a+1$ to all integers in {a,a+1,...,b} such that a becomes $a - a + 1 = 1$ and b becomes $b - a + 1$. Then we can let $n = b - a + 1$, and our possible values of $D - a + 1$ range from $1 \to n$. Now, we have n possibilities for the value of $D - a + 1$, each with equal probability $\frac{1}{n}$, so our expected value is

$$E[D-a+1] = \sum_{k=1}^{n}\frac{k}{n} = \frac{1}{n}\sum_{k=1}^{n}k = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$$

Now, we just need to calculate $E[(D-a+1)^2]$. By the definition of expected value, we know that

$$\begin{aligned}
E[(D-a+1)^2] = \quad & \sum_{k=1}^{n}\frac{k^2}{n} = \frac{1}{n}\sum_{k=1}^{n}k^2 = \frac{1}{n}\frac{n(n+1)(2n+1)}{6} \\
= \quad & \frac{2n^2+3n+1}{6}
\end{aligned}$$

Plugging this into the equation for $\sigma^2 = Var(D) = Var(D-a+1)$, we find

$$\begin{aligned}
\sigma^2 = Var(D) = Var(D-a+1) = \quad & \frac{2n^2+3n+1}{6} - \frac{n+1}{2}^2 \\
= \quad & \frac{4n^2+6n+2-3n^2-6n-3}{12} = \frac{n^2-1}{12}
\end{aligned}$$

This is true for all a and b, so we know the variance of any $D \sim DiscreteUniform(a,b)$ is

$$\sigma^2 = Var(D) = \frac{(b-a+1)^2-1}{12}$$

Assignment 9

Math 407 (Swanson) – Spring 2023
Homework 1
Due Friday 1/13, 11:59pm

Name: Emerson Kahle Section: 39981

- You must upload your solutions to Gradescope as **one single, high-quality PDF**. You can convert paper-based work to a high-quality PDF using a scanning app for mobile devices, such as Adobe Scan (free, available for iOS and Android, can do multiple pages) or many others. If necessary, you can combine or merge multiple PDF's into a single PDF using a variety of services, such as Adobe Acrobat's cloud-based merge tool.
- After you upload, you must match each question with its corresponding page using Gradescope's interface. This allows graders to spend more time giving you feedback instead of hunting through submissions.
- Answers without supporting work will receive no credit. Show your work.
- You are encouraged to work together on homework, but **you must write up your solutions separately in your own words**. Copying from your fellow students or other sources is a serious academic integrity violation. In particular, you may not use "tutoring" services which simply provide answers.
- You are encouraged to typeset your solutions in LaTeX. Source code has been provided on Blackboard. Overleaf is a popular cloud-based editor.
- Problem numbers refer to the course textbook, though the problems may have been modified significantly.

1. Fill in the final column of the table in HW 8, question 7. That is, give explicit formulas for the variances of binomial, geometric, negative binomial, Poisson, and discrete uniform random variables, using probability generating function arguments.

Solution. The completed table from HW 8 is as follows:

Distribution	Parameter(s)	PMF	PGF	μ	σ^2 / $Variance$
Binomial	n, p	$\binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$	$(pt + (1-p))^n$	np	$np(1-p)$
Geometric	p	$(1-p)^{k-1}p$	$\frac{pt}{1-(1-p)t}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative binomial	r, p	$\binom{k-1}{r-1} \cdot (1-p)^{k-r} \cdot p^r$	$(\frac{pt}{1-(1-p)t})^r$	$\frac{r}{p}$	$\frac{(1-p)r}{p^2}$
Poisson	λ	$e^{-\lambda}\frac{\lambda^k}{k!}$	$e^{(t-1)\lambda}$	λ	λ
Discrete uniform	a, b	$\frac{1}{b-a+1}$	$\frac{t^a}{b-a+1} \cdot \frac{1-t^{b-a+1}}{1-t}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$

Now, we will show the probability generating function arguments for the rightmost column in the table. We use the fact that

$$Var(X) = G"(1) + G'(1)(1 - G'(1))$$

for each distribution.

Binomial: Suppose $X \sim Binomial(n, p)$. Then $G_X(t) = (pt + (1-p))^n$, so

$$G'_X(t) = n(pt + (1-p))^{n-1}p$$

and

$$G"_X(t) = n(n-1)(pt + (1-p))^{n-2}p^2$$

so

$$\begin{aligned} Var(X) = &\quad n(n-1)(p + (1-p))^{n-2}p^2 + n(p + (1-p))^{n-1}p(1 - n(pt + (1-p))^{n-1}p) \\ = &\quad n(n-1)p^2 + np(1-np) = n^2p^2 - np^2 + np - n^2p^2 = np - np^2 = np(1-p) \end{aligned}$$

Geometric: Suppose $Y \sim Geometric(p)$. Then $G_Y(t) = \frac{pt}{1-(1-p)t}$, so

$$G'_Y(t) = p \cdot \frac{(1-(1-p)t) - (p-1)t}{(1-(1-p)t)^2} = \frac{p}{(1-(1-p)t)^2}$$

and

$$G"_Y(t) = p \cdot -2(1-(1-p)t)^{-3}(p-1) = \frac{-2p(p-1)}{(1-(1-p)t)^3}$$

so

$$\begin{aligned} Var(Y) = & \frac{-2p(p-1)}{(1-(1-p))^3} + \frac{p}{(1-(1-p))^2}(1 - \frac{p}{(1-(1-p))^2}) = \frac{-2p(p-1)}{p^3} + \frac{p}{p^2}(1-\frac{p}{p^2}) \\ = & \frac{-2}{p} + \frac{2}{p^2} + \frac{1}{p}(1-\frac{1}{p}) = \frac{-2}{p} + \frac{2}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1}{p^2} - \frac{1}{p} = \frac{1}{p^2} - \frac{p}{p^2} = \frac{1-p}{p^2} \end{aligned}$$

Negative Binomial: Suppose $N \sim NegativeBinomial(r,p)$. Then $G_N(t) = \left(\frac{pt}{1-(1-p)t}\right)^r$, so

$$G'_N(t) = \frac{rp(\frac{pt}{1-(1-p)t})^{r-1}}{(1-(1-p)t)^2}$$

which means

$$G'_N(1) = \frac{rp(\frac{p}{p})^{r-1}}{p^2} = \frac{rp}{p^2} = \frac{r}{p}$$

and

$$\begin{aligned} G"_N(t) = & \frac{r(r-1)p\left(\frac{pt}{1-(1-p)t}\right)^{r-2}\frac{p}{(1-(1-p)t)^2}(1-(1-p)t)^2 - rp(\frac{pt}{1-(1-p)t})^{r-1}\cdot 2(1-(1-p)t)(p-1)}{(1-(1-p)t)^4} \\ = & \frac{r(r-1)p^2\left(\frac{pt}{1-(1-p)t}\right)^{r-2} - rp(\frac{pt}{1-(1-p)t})^{r-1}\cdot 2(1-(1-p)t)(p-1)}{(1-(1-p)t)^4} \end{aligned}$$

which means

$$\begin{aligned} G"_N(1) = & \frac{r(r-1)p^2\left(\frac{p}{p}\right)^{r-2} - rp(\frac{p}{p})^{r-1}\cdot 2(p)(p-1)}{(p)^4} = \frac{r(r-1)p^2 - 2rp^2(p-1)}{p^4} \\ = & \frac{r(r-1) - 2r(p-1)}{p^2} = \frac{r^2 - r - 2rp + 2r}{p^2} = \frac{r^2 + r - 2rp}{p^2} = \frac{r^2 + r(1-2p)}{p^2} \end{aligned}$$

so

$$\begin{aligned} Var(N) = & \frac{r^2 + r(1-2p)}{p^2} + \frac{r}{p}(1-\frac{r}{p}) = \frac{r^2}{p^2} + \frac{r}{p^2} - \frac{2r}{p} + \frac{r}{p} - \frac{r^2}{p^2} = \frac{r}{p^2} - \frac{2r}{p} + \frac{r}{p} = \frac{r - 2rp + rp}{p^2} \\ = & \frac{r-rp}{p^2} = \frac{r(1-p)}{p^2} \end{aligned}$$

Poisson: If $P \sim Poisson(\lambda)$, then $G_P(t) = e^{(t-1)\lambda}$, so

$$G'_P(t) = \lambda e^{(t-1)\lambda}$$

and

$$G"_P(t) = \lambda^2 e^{(t-1)\lambda}$$

so

$$Var(P) = \lambda^2 + \lambda(1-\lambda) = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Discrete Uniform: IF $D \sim DiscreteUniform(a,b)$, then $G_D(t) = \frac{t^a}{b-a+1}\cdot\frac{1-t^{b-a+1}}{1-t}$. Since $Var(D) = Var(D+k)$ for any constant k, we can let $k = -a$ and the values of $D+k$ range from $a-a=0$ to $b-a$. Let $n = b-a$, and then $(D+k) = (D-a) \sim DiscreteUniform(0,n)$. Now, we can calculate $G'_{D-a}(t)$ and $G"_{D-a}(t)$ to compute $Var(D-a) = Var(D)$. For $D-a$, we have the probability generating function

$$G_{D-a}(t) = \frac{t^0}{n+1}\frac{1-t^{n+1}}{1-t} = \frac{1}{n+1}\frac{1-t^{n+1}}{1-t}$$

Therefore,

$$G'_{D_a}(t) = \frac{1}{n+1}\frac{-(n+1)t^{n+1}(1-t)+(1-t^{n+1})}{(1-t)^2}$$

which means

$$\begin{aligned} G'_{D-a}(1) = & \quad \frac{1}{n+1}\lim_{t\to 1}\frac{-(n+1)t^{n}(1-t)+(1-t^{n+1})}{(1-t)^2} \\ = & \quad \frac{1}{n+1}\lim_{t\to 1}\frac{-(n+1)nt^{n-1}(1-t)+(n+1)t^{n}-(n+1)t^{n}}{-2(1-t)} \\ = & \quad \frac{1}{n+1}\lim_{t\to 1}\frac{-(n+1)n(n-1)t^{n-2}(1-t)+(n+1)nt^{n-1}}{2} = \frac{1}{n+1}\frac{n(n+1)}{2} = \frac{n}{2} \end{aligned}$$

and

$$\begin{aligned} &G"_{D-a}(t) = \\ &\frac{-(n+1)nt^{n-1}(1-t)+(n+1)t^{n}-(n+1)t^{n})(1-t)^2+2(1-t)(-(n+1)t^{n}(1-t)+1-t^{n+1})}{(n+1)(1-t)^4} \\ &= \frac{-(n+1)nt^{n-1}(1-t)^3+2(1-t)(-(n+1)t^{n}(1-t)+1-t^{n+1})}{(n+1)(1-t)^4} \end{aligned}$$

Applying L'Hopital's Rule 4 times, we find that

$$\begin{aligned} G"_{D-a}(1) = & \quad \lim_{t\to 1}\frac{-(n+1)nt^{n-1}(1-t)^3+2(1-t)(-(n+1)t^{n}(1-t)+1-t^{n+1})}{(n+1)(1-t)^4} \\ = & \quad \frac{8n(n^2-1)}{24(n+1)} = \frac{n(n-1)}{3} = \frac{n^2-n}{3} \end{aligned}$$

Plugging $G"_{D-a}(1)$ and $G'_{D-a}(1)$ into the equation for variance, we find

$$Var(D-a) = \frac{n^2-n}{3}+\frac{n}{2}(1-\frac{n}{2}) = \frac{4n^2-4n}{12}+\frac{6n}{12}-\frac{3n^2}{12} = \frac{n^2+2n}{12} = \frac{(n+1)^2-1}{12}$$

Since $Var(D) = Var(D-a)$, we know

$$Var(D) = Var(D-a) = \frac{(n+1)^2-1}{12}$$

Since $n = b-a$, $n+1 = b-a+1$, so

$$Var(D) = \frac{(b-a+1)^2-1}{12}$$

2. (Ross P4.39) If $E[X] = 1$ and $\text{Var}(X) = 5$, find

   (a) $E[(2+X)^2]$

   (b) $\text{Var}(4+3X)$

   Solution.

   (a) Note that $E[cX] = cE[x]$ and $E[X+Y] = E[X] + E[Y]$. We can apply these two properties of expectation directly to find

$$\begin{aligned} E[(2+X)^2] = &\quad E[(X^2+4X+4)] = E[(X^2)] + E[4X] + E[4] = E[X^2] + 4E[X] + 4 \\ = &\quad E[X^2] + 4 + 4 = E[X^2] + 8 \qquad (1) \end{aligned}$$

   so we just need to compute $E[X^2]$. By definition, $Var(X) = E[X^2] - E[X]^2$, and we are given $Var(X) = 5$, so we can easily compute that

$$E[X^2] = Var(X) + E[X]^2 = 5 + 1^2 = 5 + 1 = 6$$

   Plugging 6 in for $E[X^2]$ in (1), we find

$$E[(2+X)^2] = 6 + 8 = 14$$

   (b) Note that $Var(c+X) = Var(X)$ and $Var(cX) = c^2 Var(X)$. We can apply these two properties of variance directly to find

$$Var(4+3X) = Var(3x) = 3^2 Var(X) = 9Var(X) \qquad (2)$$

   We are given that $Var(X) = 5$, so we can plug this value into (2) to find

$$Var(4+3X) = 9 \cdot 5 = 45$$

3. (a) Suppose X and Y are independent discrete random variables. Show that

$$E[XY] = E[X]E[Y].$$

(b) Give an explicit example of random variables for which $E[XY] \neq E[X]E[Y]$.

Solution.

(a) Suppose K is the set of possible values of XY. Then by the definition of expected value,

$$E[XY] = \sum_{k \in K} k\mathbb{P}(k) \qquad (1)$$

Suppose I is the set of possible values of X and J is the set of possible values of Y. Then

$$XY = k \implies X = i \in I, Y = j \in J \quad s.t. \quad i \cdot j = k$$

so

$$\mathbb{P}(XY = k) = \mathbb{P}(X = i \in I, Y = j \in J \quad s.t. \quad i \cdot j = k)$$

There could be multiple combinations, (i_1, j_1), (i_2, j_2), ... such that $i_a \cdot j_a = k$, and all outcomes are mutually disjoint (since if $X = a$, we know $X \neq b$ for all $b \neq a$, and the same holds for Y). Therefore, we know that

$$\mathbb{P}(XY = k) = \sum_{i \in I, j \in J s.t. i \cdot j = k} \mathbb{P}(X = i, Y = j)$$

We can now rewrite (1) in terms of i and j:

$$E[XY] = \sum_{k \in K} k \sum_{i \in I, j \in J s.t. i \cdot j = k} \mathbb{P}(X = i, Y = j) \qquad (2)$$

Since we are summing over all possible values of k, we are also summing over all possible combinations of i and j, and we know $k = i \cdot j$ for each combination, so we can rewrite (2) as

$$E[XY] = \sum_{i \in I, j \in J} i \cdot j\mathbb{P}(X = i, Y = j) \qquad (3)$$

Since X and Y are independent, we know $\mathbb{P}(X = i, Y = j) = \mathbb{P}(X = i)\mathbb{P}(Y = j)$ for all $i \in I$, $j \in J$, so we can rewrite (3) as

$$\begin{aligned} E[XY] = & \sum_{i \in I, j \in J} i \cdot j\mathbb{P}(X = i)\mathbb{P}(Y = j) = \sum_{i \in I} \sum_{j \in J} i \cdot j\mathbb{P}(X = i)\mathbb{P}(Y = j) \\ = & \big(\sum_{i \in I} i\mathbb{P}(X = i)\big)\big(\sum_{j \in J} j\mathbb{P}(Y = j)\big) = E[X]E[Y] \end{aligned}$$

with the last equality following from the definition of expected value. This concludes the proof that if X and Y are independent discrete random variables, $E[XY] = E[X]E[Y]$.

(b) Let $X =$ the value of a single roll of a fair six-sided die. Let $Y = X - 1$.
Then $X \sim DiscreteUniform(1, 6)$ and $Y \sim DiscreteUniform(0, 5)$, so

$$E[X] = \frac{6 + 1}{2} = \frac{7}{2} = 3.5$$

and

$$E[Y] = \frac{5 + 0}{2} = \frac{5}{2} = 2.5$$

so

$$E[X]E[Y] = \frac{7}{2}\frac{5}{2} = \frac{35}{4} = 8.75$$

Suppose K is the set of all possible values of XY. Then by the definition of expected value,

$$E(XY) = \sum_{k \in K} k\mathbb{P}(XY = k) \qquad (4)$$

If $X = 1$, $Y = 1 - 1 = 0$, so $XY = 0 \in K$.
If $X = 2$, $Y = 2 - 1 = 1$, so $XY = 2(1) = 2 \in K$.
If $X = 3$, $Y = 3 - 1 = 2$, so $XY = 3(2) = 6 \in K$.
If $X = 4$, $Y = 4 - 1 = 3$, so $XY = 4(3) = 12 \in K$.
If $X = 5$, $Y = 5 - 1 = 4$, so $XY = 5(4) = 20 \in K$.
If $X = 6$, $Y = 6 - 1 = 5$, so $XY = 6(5) = 30 \in K$.
Since X only takes values from 1 to 6, and Y's value depends entirely on the value of X, this means

$$K = \{0, 2, 6, 12, 20, 30\}$$

Furthermore, since $X \sim DiscreteUniform(1, 6)$, we know that

$$\mathbb{P}(X = i) = \frac{1}{6 - 1 + 1} = \frac{1}{6}$$

for all $1 \leq i \leq 6$.
Note: $X = i \iff XY = i(i - 1)$ for all $1 \leq i \leq 6$, so

$$\mathbb{P}(XY = k) = \mathbb{P}(XY = i(i - 1)) = \mathbb{P}(X = i) = \frac{1}{6}$$

for all $1 \leq i \leq 6$ and all $k \in K$. Plugging in $\frac{1}{6}$ for $\mathbb{P}(XY = k)$ in (4), we find

$$E[XY] = \sum_{k \in K} k\frac{1}{6} = \frac{1}{6}\sum_{k \in K} k = \frac{0 + 2 + 6 + 12 + 20 + 30}{6} = \frac{70}{6} = \frac{35}{3} \neq \frac{35}{4} = E[X]E[Y]$$

Thus, X and Y serve as an explicit example of random variables for which $E[XY] \neq E[X]E[Y]$

4. (a) Suppose X and Y are independent discrete random variables. Show that

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y).$$

(b) Give an explicit example of random variables for which $\text{Var}(X+Y) \neq \text{Var}(X) + \text{Var}(Y)$.

Solution.

(a) By the definition of variance, we know that

$$Var(X+Y) = E[(X+Y)^2] - E[(X+Y)]^2 \qquad (1)$$

Expanding and applying linearity of expectation to (1), we find

$$\begin{aligned} Var(X+Y) = \quad & E[(X^2 + 2XY + Y^2)] - (E[X] + E[Y])^2 \\ = \quad & E[X^2] + 2E[XY] + E[Y^2] - (E[X]^2 + 2E[X]E[Y] + E[Y]^2) \\ = \quad & E[X^2] + 2E[XY] + E[Y^2] - E[X]^2 - 2E[X]E[Y] - E[Y]^2 \qquad (2) \end{aligned}$$

Since X and Y are independent, we know that

$$E[XY] = E[X]E[Y]$$

Plugging $E[X]E[Y]$ in for $E[XY]$ in (2), we find

$$\begin{aligned} Var(X+Y) = \quad & E[X^2] + 2E[X]E[Y] + E[Y^2] - E[X]^2 - 2E[X]E[Y] - E[Y]^2 \\ = \quad & E[X^2] + E[Y^2] - E[X]^2 - E[Y]^2 \\ = \quad & (E[X^2] - E[X]^2) + (E[Y^2] - E[Y^2] \\ = \quad & Var(X) + Var(Y) \end{aligned}$$

with the final equality following from the definition of variance. This concludes the proof that if X and Y are independent discrete random variables, $Var(X+Y) = Var(X) + Var(Y)$.

(b) We will use the same variables X and Y as in part (b) of question 3.
That is, $X =$ the value of a single roll of a fair six-sided die, $Y = X - 1$. Since $X \sim DiscreteUniform(1, 6)$, we know

$$Var(X) = \frac{(6-1+1)^2 - 1}{12} = \frac{35}{12}$$

and since $Y \sim DiscreteUniform(0, 5)$, we know

$$Var(Y) = \frac{(5-0+1)^2 - 1}{12} = \frac{35}{12} = Var(X)$$

Therefore, we can easily compute that

$$Var(X) + Var(Y) = \frac{35}{12} + \frac{35}{12} = \frac{70}{12} = \frac{35}{6}$$

By the definition of variance, we know

$$Var(X+Y) = E[(X+Y)^2] - E[(X+Y)]^2$$

Expanding and applying linearity of expectation, we find

$$\begin{aligned} Var(X+Y) = \quad & E[X^2 + 2XY + Y^2] - E[(X+Y)]^2 \\ = \quad & E[X^2] + 2E[XY] + E[Y^2] - E[X+Y]^2 \end{aligned}$$

Since $E[(X+Y)] = E[X] + E[Y]$, and $E[X] = \frac{7}{2}$, $E[Y] = \frac{5}{2}$, we know

$$\begin{aligned} Var(X+Y) = & \quad E[X^2] + 2E[XY] + E[Y^2] - (\frac{7}{2} + \frac{5}{2})^2 \\ = & \quad E[X^2] + 2E[XY] + E[Y^2] - 6^2 \\ = & \quad E[X^2] + 2E[XY] + E[Y^2] - 36 \qquad (3) \end{aligned}$$

In part (b) of question 3, we already calculated that $E[XY] = \frac{35}{3}$, so we just need to compute $E[X^2]$ and $E[Y^2]$.
By the definition of expected value, we know

$$E[X^2] = \sum_{k=1}^{6} k^2 \mathbb{P}(X = k)$$

Since $X \sim DiscreteUniform(1, 6)$, we know $\mathbb{P}(X = k) = \frac{1}{6-1+1} = \frac{1}{6}$ for all $1 \leq k \leq 6$, which implies that

$$E[X^2] = \sum_{k=1}^{6} k^2 \frac{1}{6} = \frac{1}{6} \sum_{k=1}^{6} k^2 = \frac{1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2}{6} = \frac{91}{6}$$

Similarly, by the definition of expected value, we know

$$E[Y^2] = \sum_{k=0}^{5} k^2 \mathbb{P}(Y = k)$$

since $Y \sim DiscreteUniform(0, 5)$, we know $\mathbb{P}(Y = k) = \frac{1}{5-0+1} = \frac{1}{6}$ for all $0 \leq k \leq 5$, which implies that

$$E[Y^2] = \sum_{k=0}^{5} k^2 \frac{1}{6} = \frac{1}{6} \sum_{k=0}^{5} k^2 = \frac{0^2 + 1^2 + 2^2 + 3^2 + 4^2 + 5^2}{6} = \frac{55}{6}$$

Plugging $E[X^2]$, $E[XY]$, and $E[Y^2]$ into (3), we find that

$$\begin{aligned} Var(X+Y) = & \quad \frac{91}{6} + 2\frac{35}{3} + \frac{55}{6} - 36 \\ = & \quad \frac{91 + 140 + 55}{6} - 36 \\ = & \quad \frac{286}{6} - \frac{216}{6} \\ = & \quad \frac{70}{6} \neq \frac{35}{6} = Var(X) + Var(Y) \end{aligned}$$

Thus, X and Y serve as an explicit example of random variables for which $Var(X+Y) \neq Var(X) + Var(Y)$.

5. (Ross P4.47) Suppose that it takes at least 9 votes from a 12-member jury to convict a defendant. Suppose also that the probability that a juror votes a guilty person innocent is 0.2, whereas the probability that the juror votes an innocent person guilty is 0.1. If each juror acts independently and if 65 percent of the defendants are guilty, find the probability that the jury renders a correct decision. What percentage of defendants are convicted?

Solution.
Let J_G = the event that the jury finds the defendant guilty.
Let $J_I = J_G^c$ = the event that the jury finds the defendant innocent.
Let I = the event that the defendant is innocent.
Let $G = I^c$ the event that the defendant is guilty.
Then $J_G \cap J_I = \emptyset$, so J_G and J_I are mutually disjoint.
Similarly, $G \cap I = \emptyset$, so G and I are mutually disjoint.
Let C = the event that the jury renders a correct decision.
We need to find $\mathbb{P}(C)$ and $\mathbb{P}(J_G)$.

Note that $C = (J_G \cap G) \cup (J_I \cap I)$. If the defendant is guilty and the jury finds the defendant guilty, then the defendant cannot be innocent, and the jury cannot find the defendant innocent. Therefore, $(J_G \cap G) \cap (J_I \cap I) = \emptyset$, so $(J_G \cap G)$ and $(J_I \cap I)$ are mutually disjoint, which means

$$\mathbb{P}(C) = \mathbb{P}((J_G \cap G) \cup (J_I \cap I)) = \mathbb{P}(J_G \cap G) + \mathbb{P}(J_I \cap I)$$

Since $\mathbb{P}(X \cap Y) = \mathbb{P}(X|Y)\mathbb{P}(Y)$, we know

$$\mathbb{P}(C) = \mathbb{P}(J_G|G)\mathbb{P}(G) + \mathbb{P}(J_I|I)\mathbb{P}(I) \qquad (1)$$

Number the jurors from 1 to 12. Let $N_i = \begin{cases} 1 \text{ if the i'th juror votes guilty} \\ 0 \text{ if the i'th juror votes innocent} \end{cases}$
and let $N = \sum_{k=1}^{12} N_i$ = the total number of jurors that vote guilty.
We are given that, if the defendant is guilty, the probability that a given juror votes the defendant innocent is

$$\mathbb{P}(N_i = 0|G) = 0.2$$

for all $1 \leq i \leq 12$. This implies the probability that a given juror votes the defendant guilty (if the defendant is actually guilty) is

$$\mathbb{P}(N_i = 1|G) = 1 - \mathbb{P}(N_i = 0|G) = 1 - 0.2 = 0.8$$

Therefore, if the defendant is guilty, we can treat N as a Binomial random variable with $p = 0.8$ and $n = 12$.
Since at least 9 of the 12 jurors must vote guilty for a conviction, we know that

$$\mathbb{P}(J_G|G) = \mathbb{P}(N \geq 9|G) = \sum_{i=9}^{12} \binom{12}{i} (0.8)^i (0.2)^{12-i}$$

Similarly, if the defendant is innocent, we are given that the probability that a given juror votes the defendant guilty is

$$\mathbb{P}(N_i = 1|I) = 0.1$$

for all $1 \leq i \leq 12$. This implies the probability that a given juror votes the defendant innocent (if the defendant is actually innocent) is

$$\mathbb{P}(N_i = 0|I) = 1 - \mathbb{P}(N_i = 1|I) = 1 - 0.1 = 0.9$$

Therefore, if the defendant is innocent, we can treat N as a Binomial random variable with $p = 0.1$ and $n = 12$.

Since at least 9 of the 12 jurors must vote guilty for a conviction, if the defendant is innocent, the jury only makes the correct decision if ≤ 8 jurors vote guilty. This implies the probability that the jury votes a defendant innocent, given that they are actually innocent, is

$$\mathbb{P}(J_I|I) = \sum_{i=0}^{8} \binom{12}{i} (0.1)^i (0.9)^{12-i}$$

We are given that

$$\mathbb{P}(G) = 0.65$$

which, since $I = G^c$, implies that

$$\mathbb{P}(I) = 1 - 0.65 = 0.35$$

Plugging these computed and given values for $\mathbb{P}(J_G|G)$, $\mathbb{P}(G)$, $\mathbb{P}(J_I|I)$, and $\mathbb{P}(I)$ into (1), we find that

$$\begin{aligned} \mathbb{P}(C) = & \quad \sum_{i=9}^{12} \binom{12}{i} (0.8)^i (0.2)^{12-i} \cdot 0.65 + \sum_{i=0}^{8} \binom{12}{i} (0.1)^i (0.9)^{12-i} \cdot 0.35 \\ \approx & \quad 0.79457 \cdot 0.65 + 0.99999983 \cdot 0.35 \approx 0.8665 = 86.65\% \end{aligned}$$

Thus, the probability that the jury renders a correct decision is $\mathbb{P}(C) \approx 0.8665 = 86.65\%$.

For $\mathbb{P}(J_G)$, we can apply the Law of Total Probability to find that

$$\mathbb{P}(J_G) = \mathbb{P}(J_G|G)\mathbb{P}(G) + \mathbb{P}(J_G|I)\mathbb{P}(I) \qquad (2)$$

We already computed $\mathbb{P}(J_G|G)\mathbb{P}(G)$, and we are given $\mathbb{P}(I)$, so we just need to compute $\mathbb{P}(J_G|I)$. To do so, we can once again treat N as a Binomial random variable with $n = 12$ and $p = 0.1$. Since at least 9 of the 12 jurors must vote guilty for the defendant to be convicted, we know that

$$\mathbb{P}(J_G|I) = \mathbb{P}(N \geq 9|I) = \sum_{i=9}^{12} \binom{12}{i} (0.1)^i (0.9)^{12-i}$$

Plugging this into (2), we find

$$\begin{aligned} \mathbb{P}(J_G) = & \quad \sum_{i=9}^{12} \binom{12}{i} (0.8)^i (0.2)^{12-i} \cdot 0.65 + \sum_{i=9}^{12} \binom{12}{i} (0.1)^i (0.9)^{12-i} \cdot 0.35 \\ \approx & \quad 0.79457 \cdot 0.65 + 000000165 \cdot 0.35 \approx 0.5165 = 51.65\% \end{aligned}$$

Thus, approximately 51.65% of defendants are convicted.

6. Consider a sequence of independent Bernoulli trials each with success probability p. Let Y be the number of trials until the rth success. Let X_1 be the number of trials until the first success, let X_2 be the number of trials after the first success until the second success, and in general let X_{i+1} be the number of trials after the ith success until the $(i+1)$st success. By definition, $Y \sim \text{NegativeBinomial}(r, p)$.

   (a) Show that $Y = X_1 + X_2 + \cdots + X_r$.

   (b) Show that $X_1, X_2, \ldots, X_r$ are i.i.d. Geometric(p) random variables.

   (c) How do the probability generating functions of Y and X_i relate to each other?

   Solution.

   (a) Claim: $Y = \sum_{k=1}^{r} X_k$ for all $r \in \mathbb{N}$.
   Proof. We induct on r
   Base Case:
   $r = 1$. It takes X_1 trials until the first success, so $Y = X_1 = \sum_{k=1}^{r} X_k$, so the claim holds for the base case.
   Inductive Hypothesis:
   Assume $Y = \sum_{k=1}^{r} X_k$ for all $1 \leq r \leq j$.
   Inductive Step:
   Consider $r = j + 1$. By the inductive hypothesis, we know it takes $\sum_{k=1}^{j} X_k$ trials until the jth success. By the definition of X_{j+1}, we know it takes $j + 1$ more trials after the jth success until the $j + 1$th success. Therefore, it takes a total of

   $$Y = \sum_{k=1}^{j} X_k + X_{j+1} = \sum_{k=1}^{j+1} X_k$$

   trials until the $(j + 1)$th success. The conclusion that

   $$Y = X_1 + X_2 + \cdots + X_r$$

   follows by induction.

   (b) Since each X_i only counts the number of trials after the $(i-1)$th success until the ith success, the set of Bernoulli trials that determines X_i is mutually disjoint from the sets of Bernoulli trials that determine X_j for all $j \neq i$. Therefore, the X_i's are all mutually independent random variables. Since each X_i counts the number of independent Bernoulli trials, each with probability p, until exactly one success occurs, each $X_i \sim \textit{Geometric}(p)$ by definition of the Geometric random variable. Since the $X_1, \cdots, X_r$ are mutually independent, we know $X_1, \cdots, X_r$ are *i.i.d. Geometric*(p) random variables. For each X_i, if $X_i = k$, there must have been $k - 1$ failure trials followed by 1 success trial. Therefore, the probability mass function for each X_i is

   $$\mathbb{P}(X_i = k) = (1-p)^{k-1}p$$

   which also shows that $X_1, X_2, \cdots X_r$ are *i.i.d Geometric*(p) random variables.

   (c) Let $G_Y(t)$ = the probability generating function of Y.
   Let $G_{X_i}(t)$ = the probability generating function of X_i.
   Since $Y \sim \textit{NegativeBinomial}(r, p)$, we know it has probability generating function

   $$G_Y(t) = \Big(\frac{pt}{1-(1-p)t}\Big)^r$$

   Since $X_1, X_2, \cdots, X_r$ are *i.i.d Geometric*(p), we know they each have probability generating function

   $$G_{X_i}(t) = \frac{pt}{1-(1-p)t}$$

Therefore,

$$G_Y(t) = (G_{X_i}(t))^r$$

so the probability generating function of Y is just the probability generating function of X_i raised to the rth power.

7. Suppose $X \sim \text{Poisson}(\lambda_1), Y \sim \text{Poisson}(\lambda_2)$ are independent. Show that $X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Since $X \sim Poisson(\lambda_1)$ and $Y \sim Poisson(\lambda_2)$, we know X has probability mass function

$$\mathbb{P}(X = k) = e^{-\lambda_1} \frac{\lambda_1^k}{k!}$$

and Y has probability mass function

$$\mathbb{P}(Y = k) = e^{-\lambda_2} \frac{\lambda_2^k}{k!}$$

We want to show that $X + Y$ has probability mass function

$$\mathbb{P}(X + Y = k) = e^{-\lambda_1 - \lambda_2} \frac{(\lambda_1 + \lambda_2)^k}{k!}$$

If $X + Y = k$, then $X = i \in \{0, 1, ..., k\}$ and $Y = k - i$. Therefore,

$$\begin{aligned} \mathbb{P}(X + Y = k) = & \quad \mathbb{P}(X = i \in \{0, 1, ..., k\}, Y = k - i) \\ = & \quad \mathbb{P}((X = 0, Y = k) \cup (X = 1, Y = k - 1) \cup \cdots \cup (X = k), (Y = 0)) \end{aligned}$$

Since $(X = i, Y = k - i)$ and $(X = j, Y = k - j)$ are mutually disjoint for all $i \neq j$, we know

$$\mathbb{P}(X + Y = k) = \sum_{i=0}^{k} \mathbb{P}(X = i, Y = k - i)$$

Since X and Y are mutually disjoint, we know

$$\mathbb{P}(X = i, Y = k - i) = \mathbb{P}(X = i)\mathbb{P}(Y = k - i)$$

for all i. This implies

$$\begin{aligned} \mathbb{P}(X + Y = k) = & \quad \sum_{i=0}^{k} \mathbb{P}(X = i)\mathbb{P}(Y = k - i) = \sum_{i=0}^{k} e^{-\lambda_1} \frac{\lambda_1^i}{i!} e^{-\lambda_2} \frac{\lambda_2^{k-i}}{(k-i)!} \\ = & \quad \sum_{i=0}^{k} e^{-\lambda_1 - \lambda_2} \frac{\lambda_1^i \lambda_2^{k-i}}{i!(k-i)!} \qquad (1) \end{aligned}$$

Note that the denominator in the fraction is the same as the denominator in $\binom{k}{i}$. This suggests we should multiply both sides of (1) by $1 = \frac{k!}{k!}$ to find

$$\begin{aligned} \mathbb{P}(X + Y = k) = & \quad 1 \cdot \mathbb{P}(X + Y = k) = \frac{k!}{k!} \sum_{i=0}^{k} e^{-\lambda_1 - \lambda_2} \frac{\lambda_1^i \lambda_2^{k-i}}{i!(k-i)!} \\ = & \quad \frac{1}{k!} \sum_{i=0}^{k} e^{-\lambda_1 - \lambda_2} \lambda_1^i \lambda_2^{k-i} \frac{k!}{i!(k-i)!} \\ = & \quad \frac{e^{-\lambda_1 - \lambda_2}}{k!} \sum_{i=0}^{k} \lambda_1^i \lambda_2^{k-i} \binom{k}{i} \end{aligned}$$

Applying the Binomial Theorem, we find

$$\mathbb{P}(X + Y = k) = \frac{e^{-\lambda_1 - \lambda_2}}{k!} (\lambda_1 + \lambda_2)^k = e^{-\lambda_1 - \lambda_2} \frac{(\lambda_1 + \lambda_2)^k}{k!}$$

which is exactly what we want to show. This concludes the proof that, if $X \sim Poisson(\lambda_1)$ and $Y \sim Poission(\lambda_2)$, then $X + Y \sim Poisson(\lambda_1 + \lambda_2)$.

8. (Ross P4.60) Suppose that the number of accidents occurring on a highway each day is a Poisson random variable with parameter $\lambda = 3$.

   (a) Find the probability that 3 or more accidents occur today.

   (b) Repeat part (a) under the assumption that at least 1 accident occurs today.

   Solution.

   (a) Let $X =$ the number of accidents occurring on a highway each day. Then $X \sim Poisson(3)$. Note that $(X \geq 3) = (X \leq 2)^c$, so

   $$\mathbb{P}(X \geq 3) = 1 - \mathbb{P}(X \leq 2) \qquad (1)$$

   Since $(X \leq 2) = (X = 0) \cup (X = 1) \cup (X = 2)$, and $(X = 0)$, $(X = 1)$, and $(X = 2)$ are all mutually disjoint, we know

   $$\mathbb{P}(X \leq 2) = \mathbb{P}((X = 0) \cup (X = 1) \cup (X = 2)) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2)$$

   Since $X \sim Poisson(3)$, we know X has probability mass function

   $$\mathbb{P}(X = k) = e^{-3}\frac{3^k}{k!}$$

   so

   $$\mathbb{P}(X = 0) = e^{-3}\frac{3^0}{0!} = e^{-3}\frac{1}{1} = e^{-3}$$

   and

   $$\mathbb{P}(X = 1) = e^{-3}\frac{3^1}{1!} = e^{-3}\frac{3}{1} = 3e^{-3}$$

   and

   $$\mathbb{P}(X = 2) = e^{-3}\frac{3^2}{2!} = e^{-3}\frac{9}{2}$$

   Therefore, we know that

   $$\mathbb{P}(X \leq 2) = e^{-3} + 3e^{-3} + \frac{9}{2}e^{-3} = \frac{2 + 6 + 9}{2}e^{-3} = \frac{17}{2}e^{-3}$$

   Plugging this into (1), we find

   $$\mathbb{P}(X \geq 3) = 1 - \frac{17}{2}e^{-3} \approx 0.5768 = 57.68\%$$

   Thus, the probability that 3 or more accidents occur today is approximately 57.68%

   (b) The probability that 3 or more accidents occur today under the assumption that at least 1 accident occurs today is

   $$\mathbb{P}((X \geq 3)|(X \geq 1)) = \frac{\mathbb{P}((X \geq 3) \cap (X \geq 1))}{\mathbb{P}(X \geq 1)} \qquad (2)$$

   Since $(X \geq 3) \cap (X \geq 1) = (X \geq 3)$, we know

   $$\mathbb{P}((X \geq 3) \cap (X \geq 1)) = \mathbb{P}((X \geq 3) = 1 - \frac{17}{2}e^{-3}$$

   Since $(X \geq 1) = (X = 0)^c$, we know

   $$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - e^{-3}$$

   Plugging these values into (2), we find

   $$\mathbb{P}((X \geq 3)|(X \geq 1)) = \frac{1 - \frac{17}{2}e^{-3}}{1 - e^{-3}} \approx 0.6070 = 60.70\%$$

   Thus, the probability that at least 3 accidents occur today, under the assumption that at least 1 accident occurs today, is approximately 60.70%.

9. Suppose X is $\mathbb{Z}$-valued and there is a constant α for which $E[Q(X)] = Q(\alpha)$ for all polynomials Q. Show that $P(X = x) = \delta_{x,\alpha}$.

Solution.
We want to show that

$$\mathbb{P}(X = x) = \begin{cases} 0 \text{ if } x \neq \alpha \\ 1 \text{ if } x = \alpha \end{cases}$$

Since $E[Q(X)] = Q(\alpha)$ for all polynomials Q, if we let $Q_1(c) = c$ and $Q_2(c) = c^2$, then we have

$$E[Q_1(X)] = E[X] = Q_1(\alpha) = \alpha$$

and

$$E[Q_2(X)] = E[X^2] = Q_2(\alpha) = \alpha^2$$

This implies that

$$Var(X) = E[X^2] - E[X]^2 = \alpha^2 - \alpha^2 = 0$$

Since $Var(X) = 0$, we know X must be a constant, so

$$X = c$$

for some $c \in \mathbb{Z}$. This implies

$$E[X] = c$$

and we are given

$$E[X] = \alpha$$

so we know $c = \alpha$, which implies

$$X = c = \alpha$$

Since $X = \alpha$, we know

$$\mathbb{P}(X = \alpha) = 1$$

and

$$\mathbb{P}(X = x) = 0$$

for all $x \neq \alpha$. This completes the proof that

$$\mathbb{P}(X = x) = \begin{cases} 0 \text{ if } x \neq \alpha \\ 1 \text{ if } x = \alpha \end{cases} = \delta_{x,\alpha}$$

10. In Python, we can simulate a fair die roll using the `random` library's `randint` function. For example, the following simulates rolling 20 dice:

```
> import random
> def roll_dice(n):
>     return [random.randint(1, 6) for i in range(n)]
> print(roll_dice(20))
[4, 6, 3, 2, 3, 6, 1, 2, 3, 1, 6, 3, 2, 2, 4, 4, 3, 1, 6, 1]
```

Recall that the true mean and variance of a six-sided die roll are $\mu = 7/2 = 3.5$ and $\sigma^2 = 35/12 = 2.9166\ldots$. The sample average

$$\overline{X} = \frac{1}{20}\sum_{i=1}^{20} X_i$$

of the preceding simulation is given by

```
> def average(L):
>     return sum(x for x in L)/float(len(L))
> print(average(roll_dice(20)))
3.15
```

This single sample average is not a particularly accurate estimate of the true mean. However, we may repeat the experiment many times and average the sample averages to increase the accuracy! The following code does this by averaging 10000 such sample averages:

```
> print(average([average(roll_dice(20)) for i in range(10000)]))
3.49593
```

Note that the result is indeed quite close to the true mean of 3.5.

(a) Modify the preceding code to estimate σ^2 by averaging

$$Y = \frac{1}{20}\sum_{i=1}^{20}(X_i - \mu)^2,$$

over 10000 samples. (Note: x^2 is written as `x**2` in Python. If you are not familiar with Python, you may wish to discuss this problem with a classmate who is.)

(b) Now estimate σ^2 using the naive sample standard deviation,

$$s^2 = \frac{1}{20}\sum_{i=1}^{20}(X_i - \overline{X})^2,$$

over 10000 samples.

(c) Finally estimate σ^2 using Bessel's corrected sample standard deviation,

$$S^2 = \frac{1}{19}\sum_{i=1}^{20}(X_i - \overline{X})^2,$$

over 10000 samples.

(d) Discuss the accuracy of the above three estimates for σ^2. Do they agree with the theoretical considerations from lecture?

Solution.

(a) The modified code is as follows:

```
> def sigma_squared(L):
    > return sum((l-3.5)**2 for l in L)/float(len(L))
> print(average([sigma_squared(roll_dice(20)) for i in range (10000)]))
2.91558
```

so the estimate for σ^2 over 10000 samples is 2.91558.
Note that this result is very close to the population variance of

$$\sigma^2 = \frac{35}{12} \approx 2.9167$$

(b) The modified code is as follows:

```
> def sample_SD(L):
    > return sum((l-average(L))**2 for l in L)/float(len(L))
> print(average([sample_SD(roll_dice(20)) for i in range(10000)]))
2.78145
```

so the estimate for σ^2 over 10000 samples using naive sample standard deviation is 2.78145.
Note that this result is significantly further from the population variance than the estimate using the population mean $\mu = 3.5$.

(c) The modified code is as follows:

```
> def bessel(L):
    > return sum((l-average(L))**2 for l in L)/float(len(L)-1)
> print(average([bessel(roll_dice(20)) for i in range(10000)]))
2.92236
```

so the estimate for σ^2 over 10000 samples using Bessel's corrected sample standard deviation is 2.92236.
Note that this result is very close to the population variance, much closer than the estimate using the naive sample standard deviation.

(d) The accuracy of the above three estimates for σ^2 do agree with the theoretical considerations from lecture.
The estimate using the population mean $\mu = 3.5$ is the most accurate of the three, as the magnitude of its error is only

$$|\frac{35}{12} - 2.91558| \approx 0.00109$$

This is to be expected, as

$$E[Y] = E[\frac{1}{20}\sum_{i=1}^{20}(X_i - \mu)^2] = \frac{1}{20}\sum_{i=1}^{20}E[(X_i - \mu)^2] = \frac{Var(X_1) + \cdots + Var(X_{20}}{20} = \frac{20\sigma^2}{20} = \sigma^2$$

. Therefore, using the population mean provides an unbiased estimator for σ^2, which explains why the estimate has such little error. The estimate using the naive sample standard deviation is the least accurate of the three, as the magnitude of its error is

$$|\frac{35}{12} - 2.78145| \approx 0.13522$$

, so it underestimated σ^2 by approximately 0.13522. This is also to be expected, as the naive sample standard deviation is a biased-down estimator which *underestimates* σ^2. This explains why the estimate using naive sample standard deviation is the least accurate and underpredicts σ^2.
The estimate using Bessel's corrected sample standard deviation is almost as accurate as the estimate using $\mu = 3.5$, as the magnitude of its error is only

$$|\frac{35}{12} - 2.92236| \approx 0.00569$$

This also aligns with the theoretical considerations from lecture, as Bessel's corrected sample standard deviation, like the estimate using μ, is an unbiased estimator of σ^2. Therefore, $E[S^2] = \sigma^2$, so, over many samples, we expect the estimate using Bessel's corrected sample standard deviation to approach σ^2. This explains why this estimate has such little error over 10000 samples.

Assignment 10

Math 407 (Swanson) – Spring 2023
Homework 1
Due Friday 1/13, 11:59pm

Name: Emerson Kahle Section: 39981

- You must upload your solutions to Gradescope as **one single, high-quality PDF**. You can convert paper-based work to a high-quality PDF using a scanning app for mobile devices, such as Adobe Scan (free, available for iOS and Android, can do multiple pages) or many others. If necessary, you can combine or merge multiple PDF's into a single PDF using a variety of services, such as Adobe Acrobat's cloud-based merge tool.
- After you upload, you must match each question with its corresponding page using Gradescope's interface. This allows graders to spend more time giving you feedback instead of hunting through submissions.
- Answers without supporting work will receive no credit. Show your work.
- You are encouraged to work together on homework, but **you must write up your solutions separately in your own words**. Copying from your fellow students or other sources is a serious academic integrity violation. In particular, you may not use "tutoring" services which simply provide answers.
- You are encouraged to typeset your solutions in LaTeX. Source code has been provided on Blackboard. Overleaf is a popular cloud-based editor.
- Problem numbers refer to the course textbook, though the problems may have been modified significantly.

1. (Ross T4.28) If X is a geometric random variable, show computationally that

$$P(X = n + k \mid X > n) = P(X = k).$$

Using the interpretation of a geometric random variable, give a verbal argument as to why the preceeding equation is true.

Solution.
First, we will prove the identity computationally.
Proof. We can apply the definition of conditional probability to find that

$$\mathbb{P}(X = n + k | X > n) = \frac{\mathbb{P}(X = n + k \cap X > n)}{\mathbb{P}(X > n)} \qquad (1)$$

Since geometric random variables only take on integer values, $X > n = X \geq n + 1$, so

$$\mathbb{P}(X = n + k | X > n) = \frac{\mathbb{P}(X = n + k \cap X \geq n + 1)}{\mathbb{P}(X \geq n + 1)}$$

If $X \geq n + 1$, $X \in \{n + 1, n + 2, ...\}$. Since $X = i$ and $X = j$ are mutually disjoint for all $n + 1 \leq i \neq j$, the probability that $X \geq n + 1$ is

$$\mathbb{P}(X \geq n + 1) = \mathbb{P}(X \in \{n + 1, n + 2, ...\}) = \sum_{i=n+1}^{\infty} \mathbb{P}(X = i)$$

Since X is a geometric variable, we know $\mathbb{P}(X = k) = (1 - p)^{k-1}p$ for all $k \in \mathbb{N}$, so

$$\begin{aligned} \mathbb{P}(X \geq n + 1) = & \sum_{i=n+1}^{\infty} (1 - p)^{i-1}p = p\sum_{j=n}^{\infty}(1 - p)^j = p\big(\sum_{j=0}^{\infty}(1 - p)^j - \sum_{j=0}^{n-1}(1 - p)^j\big) \\ = & \; p\big(\frac{1}{1 - (1 - p)} - \frac{1 - (1 - p)^n}{1 - (1 - p)}\big) = \frac{p(1 - 1 + (1 - p)^n)}{p} = (1 - p)^n \end{aligned}$$

Now, let's compute $\mathbb{P}(X = n + k \cap X > n)$. If $X = n + k$ for any $k \in \mathbb{N}$, then $X = n + k \geq n + 1 > n$, so $(X = n + k \cap X > n) = (X = n + k)$, so

$$\mathbb{P}(X = n + k \cap X > n) = \mathbb{P}(X = n + k) = p(1 - p)^{n+k-1}$$

Plugging the computed values for $\mathbb{P}(X = n + k \cap X > n)$ and $\mathbb{P}(X > n)$ into (1), we find

$$\mathbb{P}(X = n + k | X > n) = \frac{p(1 - p)^{n+k-1}}{(1 - p)^n} = p(1 - p)^{n+k-1-n} = p(1 - p)^{k-1} = \mathbb{P}(X = k)$$

for all $k \in \mathbb{N}$.
For $k \leq 0$, $\mathbb{P}(X = n + k|X > n) = 0 = \mathbb{P}(X = k)$. The first equality follows because it is impossible for $X = n + k \leq n$ if we are given that $X > n$. The second equality follows since $\mathbb{P}(X = k) = 0$ for $k \notin \mathbb{N}$ for all geometric random variables. This concludes the computational proof for all integers k.

We will now prove the identity using the interpretation of a geometric random variable X as the number of independent Bernoulli trials until the first success, where the probability of success on each trial is independently p.
This provides the interpretation that $\mathbb{P}(X = k) =$ the probability that it takes exactly k trials until the first success. We have $X = k$ if and only if the first $k - 1$ trials are all failures and the kth trial is a success.
Now, let's consider $\mathbb{P}(X = n + k|X > n)$. We can interpret this as the probability that it takes exactly $n + k$ trials until the first success, given that it takes more than n trials until the first success. We can

also interpret this as the probability that it takes k trials *after* the nth trial until the first success, given that it takes more than n trials until the first success. Since we know it takes more than n trials until the first success, we can count only the trials after the nth trial until the first success. For there to be exactly k such trials until the first success, the first $k-1$ such trials must all be failures, and the kth such trial must be a success.
Therefore, both $\mathbb{P}(X = k)$ and $\mathbb{P}(X = n + k|X > n)$ can be interpreted as the probability that, in a sequence of k independent Bernoulli trials, the first $k-1$ independent Bernoulli will result in failures and the kth will result in the first success. This concludes the verbal proof that

$$\mathbb{P}(X = n + k|X > n) = \mathbb{P}(X = k)$$

for all $k \in \mathbb{N}$ and a geometric random variable X.

2. (Ross P7.6) A fair die is rolled 10 times. Calculate the expected sum of the 10 rolls.

Solution.
Let $X_1, X_2, ..., X_{10}$ be random variables, where X_i = the value of the ith roll of the die.
Let $Y = X_1 + \cdots + X_{10} = \sum_{i=1}^{10} X_i$ = the sum of the 10 rolls of the die.
Then we want to calculate

$$E[Y] = E[\sum_{i=1}^{10} X_i]$$

Applying the fact that $E[A + B] = E[A] + E[B]$ for all random variables A and B, we find

$$E[Y] = \sum_{i=1}^{10} E[X_i]$$

Since $X_i \sim DiscreteUniform(1, 6)$ for all $1 \leq i \leq 10$, each of the 10 rolls of the die should have the same expected value. That is,

$$E[X_i] = E[X_1] = \cdots = E[X_{10}]$$

This implies that

$$E[Y] = \sum_{i=1}^{10} E[X_1] = 10E[X_1]$$

We can calculate $E[X_1]$ directly using the definition of expected value. Since $X_1 \sim DiscreteUniform(1, 6)$, we know $\mathbb{P}(X = k) = \frac{1}{6-1+1} = \frac{1}{6}$ for all $k \in \{1, 2, 3, 4, 5, 6\}$. This implies that

$$E[X_1] = \sum_k k\mathbb{P}(X = k) = \sum_{k=1}^{6} \frac{1}{6}k = \frac{1}{6}\sum_{k=1}^{6} k = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = \frac{21}{6} = \frac{7}{2} = 3.5$$

We could also note that, since $X_1 \sim DiscreteUniform(1, 6)$, it has an expected value of

$$E[X_1] = \frac{6 + 1}{2} = \frac{7}{2} = 3.5$$

Plugging in $E[X_1] = 3.5$ into the above equation for $E[Y]$, we find

$$E[Y] = 10E[X_1] = 10(3.5) = 35$$

Therefore, the expected sum of the 10 rolls of the die is 35.

3. (Ross P7.7) Suppose that A and B each randomly and independently choose 3 of 10 objects. Find the expected number of objects

   (a) chosen by both A and B;

   (b) not chosen by either A or B;

   (c) chosen by exactly one of A and B.

   Solution.

   (a) Let $X =$ the number of objects chosen by both A and B. Since A and B both only choose 3 objects, they can choose maximally 3 of the same objects. Furthermore, since there are 10 objects to choose from, it is possible that A chooses 3 objects and B chooses 3 objects from the 7 objects which A did not choose. In this case, the number of objects chosen by both A and B is 0. Thus, X can take on any value in $\{0, 1, 2, 3\}$. By the definition of expected value,

$$E[X] = \sum_{k=0}^{3} k\mathbb{P}(X = k) = \sum_{k=1}^{3} k\mathbb{P}(X = k) \qquad (1)$$

   There are $\binom{10}{3}$ ways A can choose its objects and $\binom{10}{3}$ ways B can choose its objects, for a total of

$$\binom{10}{3}\binom{10}{3} = \binom{10}{3}^2$$

   total possible ways in which A and B can select their objects, where all outcomes are equally likely. For $\mathbb{P}(X = 3)$ A can choose its 3 objects in $\binom{10}{3}$ ways, and B must choose the exact same subset of 3 objects that A chooses, which can be done in exactly 1 way. This yields $\binom{10}{3}$ total possible ways for the number of objects chosen by both A and B to be 3. Since all outcomes are equally likely, this yields

$$\mathbb{P}(X = 3) = \frac{\binom{10}{3}}{\binom{10}{3}^2} = \frac{1}{\binom{10}{3}}$$

   For $\mathbb{P}(X = 2)$, A can once again choose its 3 objects in $\binom{10}{3}$ ways, but B must now choose 2 out of the three objects chosen by A and 1 out of the 7 objects not chosen by A. There are $\binom{3}{2}\binom{7}{1}$ ways for B to do this. This yields $\binom{10}{3}\binom{3}{2}\binom{7}{1}$ total ways for the number of objects chosen by both A and B to be 2. Since all outcomes are equally likely, this yields

$$\mathbb{P}(X = 2) = \frac{\binom{10}{3}\binom{7}{1}\binom{3}{2}}{\binom{10}{3}^2} = \frac{\binom{7}{1}\binom{3}{2}}{\binom{10}{3}}$$

   For $\mathbb{P}(X = 1)$, A can once again choose its 3 objects in $\binom{10}{3}$ ways, but B must now choose 1 of those 3 objects and 2 of the 7 objects not chosen by A. B can do this in $\binom{3}{1}\binom{7}{2}$ ways. This yields $\binom{10}{3}\binom{7}{2}\binom{3}{1}$ total ways for the number of objects chosen by both A and B to be 1. Since all outcomes are equally likely, this yields

$$\mathbb{P}(X = 1) = \frac{\binom{10}{3}\binom{7}{2}\binom{3}{1}}{\binom{10}{3}^2} = \frac{\binom{7}{2}\binom{3}{1}}{\binom{10}{3}}$$

   Plugging these results into (1), we find

$$E[X] = 1 \cdot \frac{\binom{7}{2}\binom{3}{1}}{\binom{10}{3}} + 2 \cdot \frac{\binom{7}{1}\binom{3}{2}}{\binom{10}{3}} + 3 \cdot \frac{1}{\binom{10}{3}} = \frac{\binom{7}{2}\binom{3}{1} + 2\binom{7}{1}\binom{3}{2} + 3}{\binom{10}{3}} = 0.90$$

   Thus, the expected number of objects chosen by both A and B is 0.90.

Note: We could also note that the probability that any individual object is chosen by A is $\mathbb{P}($ chosen by A$) = \frac{3}{10}$ and the probability that any individual object is chosen by B is $\mathbb{P}($ chosen by B$) = \frac{3}{10}$. These probabilities are independent, so the probability that an individual object is chosen by both A and B is $\mathbb{P}($ chosen by A and $B) = \frac{3}{10}\frac{3}{10} = \frac{9}{100}$. There are 10 total objects, each with a $\frac{9}{100}$ probability of being chosen by both A and B, which yields $10 \cdot \frac{9}{100} = \frac{90}{100} = \frac{9}{10} = 0.90$ for the expected number of objects chosen by both A and B.

(b) Let $Y =$ the number of objects not chosen by either A or B. If A and B choose the same 3 objects, then the number of objects not chosen by either A or B is $10 - 3 = 7$. If A and B choose mutually disjoint sets of 3 objects, then the number of objects not chosen by A or B is $10 - 6 = 4$. Thus, Y can take on any values in $\{4, 5, 6, 7\}$. By the definition of expected value, we know

$$E[Y] = \sum_{k=4}^{7} k\mathbb{P}(Y = k) \qquad (2)$$

For $\mathbb{P}(Y = 4)$, A can pick its objects in $\binom{10}{3}$ ways, but B must choose all 3 of its objects from the 7 objects not chosen by A. There are $\binom{7}{3}$ ways for B to do this. This yields

$$\binom{10}{3}\binom{7}{3}$$

total ways for the number of objects not chosen by either A or B to be 4. Since all outcomes are equally likely, this yields

$$\mathbb{P}(Y = 4) = \frac{\binom{10}{3}\binom{7}{3}}{\binom{10}{3}^2} = \frac{\binom{7}{3}}{\binom{10}{3}}$$

For $\mathbb{P}(Y = 5)$, A can pick its objects in $\binom{10}{3}$ ways, but B must choose 2 of its objects from the 7 objects not chosen by A and 1 object from the 3 objects chosen by A. There are $\binom{7}{2}\binom{3}{1}$ ways for B to do this. This yields

$$\binom{10}{3}\binom{7}{2}\binom{3}{1}$$

total ways for the number of objects not chosen by either A or B to be 5. Since all outcomes are equally likely, this yields

$$\mathbb{P}(Y = 5) = \frac{\binom{10}{3}\binom{7}{2}\binom{3}{1}}{\binom{10}{3}^2} = \frac{\binom{7}{2}\binom{3}{1}}{\binom{10}{3}}$$

For $\mathbb{P}(Y = 6)$, A can pick its objects in $\binom{10}{3}$ ways, but B must choose 1 of its objects from the 7 objects not chosen by A and 2 objects from the 3 objects chosen by A. There are $\binom{7}{1}\binom{3}{2}$ ways for B to do this. This yields

$$\binom{10}{3}\binom{7}{1}\binom{3}{2}$$

total ways for the number of objects not chosen by either A or B to be 6. Since all outcomes are equally likely, this yields

$$\mathbb{P}(Y = 6) = \frac{\binom{10}{3}\binom{7}{1}\binom{3}{2}}{\binom{10}{3}^2} = \frac{\binom{7}{1}\binom{3}{2}}{\binom{10}{3}}$$

For $\mathbb{P}(Y = 7)$, A can pick its objects in $\binom{10}{3}$ ways, but B must choose all 3 of its objects from the 3 objects chosen by A. There is 1 way for B to do this. This yields

$$\binom{10}{3}$$

total ways for the number of objects not chosen by either A or B to be 7. Since all outcomes are equally likely, this yields

$$\mathbb{P}(Y=7)=\frac{\binom{10}{3}}{\binom{10}{3}^2}=\frac{1}{\binom{10}{3}}$$

Plugging these values into (2), we find

$$E[Y]=4\cdot\frac{\binom{7}{3}}{\binom{10}{3}}+5\cdot\frac{\binom{7}{2}\binom{3}{1}}{\binom{10}{3}}+6\cdot\frac{\binom{7}{1}\binom{3}{2}}{\binom{10}{3}}+7\cdot\frac{1}{\binom{10}{3}}=\frac{4\binom{7}{3}+5\binom{7}{2}\binom{3}{1}+6\binom{7}{1}\binom{3}{2}+7}{\binom{10}{3}}=4.9$$

Thus, the expected number of objects not chosen by either A or B is 4.9.

Note: We could also note that the probability that an individual object is not chosen by A is $\mathbb{P}(\text{ not chosen by A })=1-\mathbb{P}(\text{ chosen by A })=1-\frac{3}{10}=\frac{7}{10}$. The probability that an individual object is not chosen by B is also $\mathbb{P}(\text{ not chosen by B })=1-\mathbb{P}(\text{ chosen by B })=1-\frac{3}{10}=\frac{7}{10}$. These probabilities are independent, so $\mathbb{P}(\text{ not chosen by either } A \text{ or } B)=\frac{7}{10}\frac{7}{10}=\frac{49}{100}$. There are 10 objects in total, each with a $\frac{49}{100}$ probability of not being chosen by either A or B, which yields $10\cdot\frac{49}{100}=\frac{490}{100}=\frac{49}{10}=4.9$ for the expected number of objects not chosen by either A or B.

(c) Let $Z=$ the number of objects chosen by exactly one of A and B. If A and B select the same set of three objects, then the number of objects chosen by exactly one of A and B is 0. If 2 of the numbers are chosen by both A and B, then the number of objects chosen by exactly one of A and B is 2. If 1 of the objects is chosen by both A and B, then the number of objects chosen by exactly one of A and B is 4. If A and B select mutually disjoint sets of 3 objects, then the number of objects chosen by exactly one of A and B is 6. Thus, Z can take on any values in $I=\{0,2,4,6\}$. By the definition of expected value,

$$E[Z]=\sum_{k\in I}k\mathbb{P}(Z=k)=\sum_{k\in(I-\{0\})}k\mathbb{P}(Z=k)\qquad(3)$$

For $\mathbb{P}(Z=6)$, A can pick its objects in $\binom{10}{3}$ ways, but B must choose all 3 of its objects from the 7 objects not chosen by A. There are $\binom{7}{3}$ ways for B to do this. This yields

$$\binom{10}{3}\binom{7}{3}$$

total ways for the number of objects chosen by exactly one of A and B to be 6. Since all outcomes are equally likely, this yields

$$\mathbb{P}(Z=6)=\frac{\binom{10}{3}\binom{7}{3}}{\binom{10}{3}^2}=\frac{\binom{7}{3}}{\binom{10}{3}}$$

For $\mathbb{P}(Z=4)$, A can pick its objects in $\binom{10}{3}$ ways, but B must pick 2 objects from the 7 objects not chosen by A and 1 object from the three chosen by A. There are $\binom{7}{2}\binom{3}{1}$ ways for B to do this. This yields

$$\binom{10}{3}\binom{7}{2}\binom{3}{1}$$

total ways for the number of objects chosen by exactly one of A and B to be 4. Since all outcomes are equally likely, this yields

$$\mathbb{P}(Z=4)=\frac{\binom{10}{3}\binom{7}{2}\binom{3}{1}}{\binom{10}{3}^2}=\frac{\binom{7}{2}\binom{3}{1}}{\binom{10}{3}}$$

For $\mathbb{P}(Z=2)$, A can pick its objects in $\binom{10}{3}$ ways, but B must pick 1 objects from the 7 objects not chosen by A and 2 objects from the three chosen by A. There are $\binom{7}{1}\binom{3}{2}$ ways for B to do this. This yields

$$\binom{10}{3}\binom{7}{1}\binom{3}{2}$$

total ways for the number of objects chosen by exactly one of A and B to be 2. Since all outcomes are equally likely, this yields

$$\mathbb{P}(Z=2)=\frac{\binom{10}{3}\binom{7}{1}\binom{3}{2}}{\binom{10}{3}^2}=\frac{\binom{7}{1}\binom{3}{2}}{\binom{10}{3}}$$

Plugging these values into (3), we find

$$E[Z]=2\cdot\frac{\binom{7}{1}\binom{3}{2}}{\binom{10}{3}}+4\cdot\frac{\binom{7}{2}\binom{3}{1}}{\binom{10}{3}}+6\cdot\frac{\binom{7}{3}}{\binom{10}{3}}=\frac{2\binom{7}{1}\binom{3}{2}+4\binom{7}{2}\binom{3}{1}+6\binom{7}{3}}{\binom{10}{3}}=4.2$$

Thus, the expected number of objects chosen by exactly 1 of A and B is 4.2.

Note 1: We could also note that the probability of any individual object being chosen by A and not B is $\mathbb{P}(\text{ chosen by A but not B })=\mathbb{P}(\text{ chosen by A })\mathbb{P}(\text{ not chosen by } B\text{ })=\frac{3}{10}\frac{7}{10}=\frac{21}{100}$. Similarly, the probability of an individual object being chosen by B but not A is $\mathbb{P}(\text{ chosen by } B \text{ but not } A\text{ })=\mathbb{P}(\text{ chosen by } B\text{ })\mathbb{P}(\text{ not chosen by } A\text{ })=\frac{3}{10}\frac{7}{10}=\frac{21}{100}$. Therefore, the probability of an individual object being chosen by exactly one of A and B is $\mathbb{P}(\text{ chosen by A but not B })+\mathbb{P}(\text{ chosen by B but not A})=\frac{21}{100}+\frac{21}{100}=\frac{42}{100}$. This is true for all objects, and there are 10 objects in total, which yields $10\cdot\frac{42}{100}=\frac{420}{100}=\frac{42}{10}=4.2$ for the expected number of objects chosen by exactly one of A and B.

Note 2: We can also note that the expected number of total objects is 10. Any object must be chosen by both A and B, be chosen by exactly one of A and B, or not be chosen by either A or B. Therefore, $10=E[X]+E[Y]+E[Z]$. From parts (a) and (b), we calculated that $E[X]=0.90$ and $E[Y]=4.90$. This implies that the expected number of objects chosen by exactly one of A and B is $E[Z]=10-E[X]-E[Y]=10-4.90-0.90=10-5.80=4.20$.

4. (Ross P7.18) Cards from an ordinary deck of 52 playing cards are turned face up one at a time. If the 1st card is an ace, or the 2nd is a deuce, or the 3rd a three, or ..., or the 13th a king, or the 14th an ace, and so on, we say that a match occurs. Note that we do not require that the $(12n+1)$ card be any particular ace for a match to occur but only that it be an ace. Compute the expected number of matches that occur.

Solution.

Let $X_1, ..., X_{52}$ be random variables where $X_i = \begin{cases} 1 \text{ if the } i\text{th card is a match} \\ 0 \text{ otherwise.} \end{cases}$

Let $Y = X_1 + \cdots + X_{52} = \sum_{i=1}^{52} X_i$. Then $Y =$ the number of matches that occur in the entire 52 card deck. We want to compute

$$E[Y] = E[X_1 + \cdots + X_{52}] = E[\sum_{i=1}^{52} X_i]$$

Applying linearity of expectation, we find

$$E[Y] = \sum_{i=1}^{52} E[X_i] \qquad (1)$$

Consider the ith card in the deck. It could be any of the 52 cards in the deck. Of these 52 cards, exactly 4 of them result in the ith card having a match. Assuming the cards in the deck are arranged randomly, this yields $\mathbb{P}(X_i = 1) = \frac{4}{52} = \frac{1}{13}$ for all $1 \leq i \leq 52$. Applying the definition of expected value, we find

$$E[X_i] = \sum_{k=0}^{1} k\mathbb{P}(X_i = k) = \sum_{k=1}^{1} k\mathbb{P}(X_i = k) = \mathbb{P}(X_i = 1) = \frac{1}{13}$$

for all $1 \leq i \leq 52$. Plugging $\frac{1}{13}$ in for $E[X_i]$ in (1), we find

$$E[Y] = \sum_{i=1}^{52} \frac{1}{13} = 52 \cdot \frac{1}{13} = \frac{52}{13} = 4$$

Thus, the expected number of matches that occur in the entire deck is 4.

5. (Ross T5.7) The standard deviation of X, denoted $\mathrm{SD}(X)$, is given by

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}.$$

Find $\mathrm{SD}(aX + b)$ if X has variance σ^2.

Solution.
By the definition of standard deviation, we know

$$SD(aX + b) = \sqrt{Var(aX + b)} \qquad (1)$$

Claim: $Var(X + b) = Var(X)$ for all constants b.
Proof.
By the definition of variance,

$$\begin{aligned} Var(X + b) = &\quad E[(X + b)^2] - E[X + b]^2 = E[X^2 + 2bX + b^2] - (E[X] + b)^2 \\ = &\quad E[X^2] + 2bE[X] + E[b^2] - (E[X]^2 + 2bE[X] + b^2) \\ = &\quad E[X^2] - E[X]^2 + (2bE[X] - 2bE[X]) + (b^2 - b^2) = E[X^2] - E[X]^2 = Var(X) \end{aligned}$$

This directly implies that

$$Var(aX + b) = Var(aX)$$

We know from lecture that $Var(cX) = c^2 Var(X)$ for all constants c, which implies

$$Var(aX + b) = Var(aX) = a^2 Var(X) = a^2\sigma^2$$

since X has variance $Var(X) = \sigma^2$. Plugging $a^2\sigma^2$ for $Var(aX + b)$ in (1), we find

$$SD(aX + b) = \sqrt{Var(aX + b)} = \sqrt{a^2\sigma^2} = |a\sigma|$$

Thus, if a random variable X has variance σ^2, then the standard deviation of $aX+b$ is $SD(aX+b) = |a\sigma|$.

6. Recall that the ("raw") *moments* of a random variable X are $E[X^k]$ for $k = 1, 2, 3, \ldots$. Write $\mu_k = E[X^k]$.

   (a) Describe the first two moments (i.e. μ_1, μ_2) qualitatively. If X is measured in "meters", what are the units of μ_k?

   (b) The *central moments* of X are defined by $\alpha_k = E[(X - E[X])^k]$. Describe the first two central moments (i.e. α_1, α_2) qualitatively.

   (c) Express μ_4 in terms of $\mu, \alpha_1, \alpha_2, \alpha_3, \alpha_4$.

   Solution.

   (a) The first raw moment, $\mu_1 = E[X^1] = E[X]$ is simply the expected value of X. This can be thought of as a weighted average of all possible values that X can take on, where the weights are assigned based on the probabilities that X equals each individual value. In other words, $\mu_1 = E[X]$ describes the raw mean of a random variable X. It also describes the average raw distance from the origin of a random variable X.
   The second raw moment, $\mu_2 = E[X^2]$ is the expected value of X^2. This can be thought of as a weighted average of the squares of all possible values that X can take on, where the weights are again assigned based on the probabilities that X equals each individual value. In other words, $\mu_2 = E[X^2]$ describes the average squared distance from the origin of a random variable X.
   If X is measured in "meters", then the units of μ_k are " meters^k ". For example, the units of μ_1 are "meters", the units of μ_2 are "square meters" (" meters^2 "), and so on.

   (b) The first central moment, $\alpha_1 = E[(X-E[X])^1] = E[(X-E[X])]$, is the expected difference between X and the expected value of X. This can be thought of as a weighted average of the raw differences between each possible value of X and the mean of X, where the weights are assigned based on the probabilities that X equals each individual value. The first central moment $\alpha_1 = E[X - E[X]]$ is always zero because the weighted average of the negative differences must have the same magnitude as the weighted average of the positive differences for $E[X]$ to be the weighted average of all possible values of X. This makes sense, as X is expected to be $E[X]$, so the expected difference between X and $E[X]$ should be 0.
   The second central moment,

$$\begin{aligned} E[(X - E[X])^2] = & \quad E[(X - E[X])^2] = E[X^2 - 2XE[X] + E[X^2]] \\ = & \quad E[X^2 - 2E[X]^2 + E[X]^2) = E[X^2] - E[X]^2 \end{aligned}$$

   can be easily seen to equal the variance $Var(X)$. This can be interpreted as the weighted average of the squared differences between each possible value of X and the mean of X, where the weights are again assigned based on the probabilities that X equals each possible value. In other words, $\alpha_2 = E[(X - E[X])^2]$ relates the variability of X by describing the expected squared magnitude of the absolute distance between X and the mean of X. Since negative raw differences between X and $E[X]$ still have positive squared magnitudes, α_2 should always be positive for any X that has nonzero probabilities of equalling multiple different values.

   (c) First, let's write simplify α_1, α_2, α_3, and α_4: For α_1, we have

$$\alpha_1 = E[(X - E[X])] = E[X] - E[X] = 0 \qquad (1)$$

   For α_2, we have

$$\alpha_2 = E[(X - E[X])^2] = E[X^2 - 2XE[X] + E[X]^2] = E[X^2] - E[X]^2 = E[X^2] - \mu^2 \qquad (2)$$

   For α_3, we have

$$\begin{aligned} \alpha_3 = & \quad E[(X - E[X])^3] = E[(X^3 - 3X^2E[X] + 3XE[X]^2 - E[X]^3] \\ = & \quad E[X^3] - 3E[X]E[X^2] + 3E[X]^2E[X] - E[X]^3 \\ = & \quad E[X^3] - 3\mu E[X^2] + 3E[X]^3 - E[X]^3 \\ = & \quad E[X^3] - 3\mu E[X^2] + 2\mu^3 \qquad (3) \end{aligned}$$

For α_4, we have

$$\begin{aligned}\alpha_4 = \quad & E[(X - E[X])^4] = E[(X^4 - 4X^3E[X] + 6X^2E[X]^2 - 4XE[X]^3 + E[X]^4)] \\ = \quad & E[X^4] - 4E[X]E[X^3] + 6E[X]^2E[X^2] - 4E[X]^3E[X] + E[X]^4 \\ = \quad & E[X^4] - 4\mu E[X^3] + 6\mu^2 E[X^2] - 3\mu^4 \qquad (4)\end{aligned}$$

Note that μ_4 is the first term in the sum equivalent to α_4. This suggests we should modify α_4 in some way to find μ_4. First, let's cancel the second term in the sum equivalent to α_4, which is $-4\mu E[X^3]$. We can do this via adding 4μ copies of α_3 to α_4:

$$\begin{aligned}\alpha_4 + 4\mu\alpha_3 = \quad & E[X^4] - 4\mu E[X^3] + 6\mu^2 E[X^2] - 3\mu^4 + 4\mu\big(E[X^3] - 3\mu E[X^2] + 2\mu^3\big) \\ = \quad & E[X^4] + (-4\mu E[X^3] + 4\mu E[X^3]) + (6\mu^2 E[X^2] - 12\mu^2 E[X^2]) + (-3\mu^4 + 8\mu^4) \\ = \quad & E[X^4] - 6\mu^2 E[X^2] + 5\mu^4\end{aligned}$$

Now we can cancel the $-6\mu^2 E[X^2]$ term by adding $6\mu^2$ copies of α_2 to $\alpha_4 + 4\mu\alpha_3$:

$$\begin{aligned}\alpha_4 + 4\mu\alpha_3 + 6\mu^2\alpha_2 = \quad & E[X^4] - 6\mu^2 E[X^2] + 5\mu^4 + 6\mu^2\big(E[X^2] - \mu^2\big) \\ = \quad & E[X^4] + (-6\mu^2 E[X^2] + 6\mu^2 E[X^2]) + (5\mu^4 - 6\mu^4) \\ = \quad & E[X^4] - \mu^4\end{aligned}$$

Now, we can just add μ^4 to find

$$\alpha_4 + 4\mu\alpha_3 + 6\mu^2\alpha_2 + \mu^4 = E[X^4] = \mu_4$$

Since $\alpha_1 = 0$, we can add α_1 to express $\mu_4 = E[X^4]$ in terms of μ, α_1, α_2, α_3, and α_4 as follows:

$$\mu_4 = E[X^4] = \alpha_4 + 4\mu\alpha_3 + 6\mu^2\alpha_2 + \alpha_1 + \mu^4$$

7. Let X be a random variable and let $Z = (X - \mu)/\sigma$ be the standardized random variable associated to X. Write $\mu_{X,k}$ or $\mu_{Z,k}$ and $\alpha_{X,k}$ or $\alpha_{Z,k}$ for the moments and central moments of X or Z.

   (a) How are $\mu_{Z,k}$ and $\alpha_{Z,k}$ related?

   (b) How are $\mu_{Z,k}$ and $\alpha_{X,k}$ related?

   (c) The "higher moments" $\mu_3, \mu_4, \ldots$ often appear in more refined calculations (e.g. in the famous Berry–Esseen theorem). Compute the higher moments of an exponential random variable.

Solution.

(a) We know from lecture that the standardized random variable $Z = \frac{X-\mu}{\sigma}$ has mean $\mu_Z = E[Z] = 0$. By definition, $\mu_{Z,k}$ is

$$\mu_{Z,k} = E[Z^k]$$

and $\alpha_{Z,k}$ is

$$\alpha_{Z,k} = E[(Z - E[Z])^k]$$

Since $E[Z] = \mu_Z = 0$, we know that

$$\alpha_{Z,k} = E[(Z-0)^k] = E[Z^k] = \mu_{Z,k}$$

Thus, $\mu_{Z,k}$ and $\alpha_{Z,k}$ are equivalent.

(b) By definition, $\alpha_{X,k}$ is

$$\alpha_{X,k} = E[(X-\mu)^k]$$

Since $Z = \frac{X-\mu}{\sigma}$, we can write $\mu_{Z,k}$ as

$$\mu_{Z,k} = E[Z^k] = E[\big(\frac{X-\mu}{\sigma}\big)^k] = E[\frac{(X-\mu)^k}{\sigma^k}]$$

Since $\frac{1}{\sigma^k}$ is a constant, and $E[aX] = aE[X]$ for all constants a and random variables X, we have

$$\mu_{Z,k} = \frac{1}{\sigma^k}E[(X-\mu)^k] = \frac{1}{\sigma^k}\alpha_{X,k} = \frac{\alpha_{X,k}}{\sigma^k}$$

Thus, $\mu_{Z,k}$ and $\alpha_{X,k}$ are related in that dividing $\alpha_{X,k}$ by σ^k yields $\mu_{Z,k}$.

(c) Let $X \sim Exponential(\lambda)$. By the definition of an exponential random variable, X has the probability density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \text{ if } x > 0 \\ 0 \text{ otherwise.} \end{cases}$$

We want to find a formula for $E[X^k]$ for all $k \in \mathbb{N}$. We will do so by first finding a recurrence relation for $E[X^k]$, then proving the closed form of the recurrence relation using induction.
First, we will find the recurrence relation for $E[X^k]$.
By the definition of expected value of a continuous random variable,

$$E[X^k] = \int_{-\infty}^{\infty} x^k f(x)dx$$

for all $k \in \mathbb{N}$. The base case of our recurrence relation will occur when $k \in \mathbb{N}$ is minimal (i.e. when $k = 1$).
When $k = 1$, we can compute directly that

$$\begin{aligned} E[X^k] = E[X] = & \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{0} x \cdot 0 dx + \int_{0}^{\infty} x\lambda e^{-\lambda x}dx \\ = & \int_{0}^{\infty} x\lambda e^{-\lambda x}dx = \lambda \int_{0}^{\infty} xe^{-\lambda x}dx \end{aligned}$$

Now, we can apply integration by parts with $u = x$, $du = dx$, $dv = e^{-\lambda x}dx$, and $v = \frac{e^{-\lambda x}}{-\lambda}$ to find

$$E[X] = \lambda\big(\frac{xe^{-\lambda x}}{-\lambda}\Big|_0^\infty + \frac{1}{\lambda}\int_0^\infty e^{-\lambda x}dx\big) = -xe^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x}dx \qquad (1)$$

We can evaluate the leftmost term in the sum from (1) as

$$-xe^{-\lambda x}\Big|_0^\infty = \lim_{x\to\infty} -xe^{-\lambda x} + 0 \cdot e^{-\lambda\cdot 0} = \lim_{x\to\infty}\frac{-x}{e^{\lambda x}} + 0 = \lim_{x\to\infty}\frac{-x}{e^{\lambda x}}$$

Applying L'Hopital's Rule once, we find

$$-xe^{-\lambda x}\Big|_0^\infty = \lim_{x\to\infty}\frac{-1}{\lambda e^{\lambda x}} = 0$$

We can evaluate the rightmost term in the sum from (1) directly as

$$\int_0^\infty e^{-\lambda x}dx = \frac{e^{-\lambda x}}{-\lambda}\Big|_0^\infty = \lim_{x\to\infty}\frac{e^{-\lambda x}}{-\lambda} + \frac{e^{-\lambda\cdot 0}}{\lambda} = 0 + \frac{1}{\lambda} = \frac{1}{\lambda}$$

Plugging these values into (1), we find

$$E[X^k] = E[X] = 0 + \frac{1}{\lambda} = \frac{1}{\lambda}$$

when $k = 1$. This serves as the base case for our recurrence relation.
Now, we need to express $E[X^k]$ recursively for all $k \in \mathbb{N}$ s.t. $k \geq 2$. We can apply the definition of expected value of a continuous random variable to find

$$\begin{aligned} E[X^k] = & \int_{-\infty}^\infty x^k f(x)dx = \lambda\big(\int_{-\infty}^0 x^k \cdot 0dx + \int_0^\infty x^k \cdot \lambda e^{-\lambda x}dx\big) \\ = & \lambda(0 + \int_0^\infty x^k \cdot \lambda e^{-\lambda x}dx) = \lambda\int_0^\infty x^k e^{-\lambda x}dx \end{aligned}$$

We can apply integration by parts with $u = x^k$, $du = kx^{k-1}dx$, $dv = e^{-\lambda x}dx$, and $v = \frac{e^{-\lambda x}}{-\lambda}$ to find

$$\begin{aligned} E[X^k] = & \lambda\big(\frac{x^k e^{-\lambda x}}{-\lambda}\Big|_0^\infty + \frac{k}{\lambda}\int_0^\infty x^{k-1}e^{-\lambda x}dx\big) \\ = & -x^k e^{-\lambda x}\Big|_0^\infty + k\frac{E[X^{k-1}]}{\lambda} = -x^k e^{-\lambda x}\Big|_0^\infty + \frac{kE[X^{k-1}]}{\lambda} \end{aligned}$$

We can simplify further by evaluating the $-x^k e^{-\lambda x}\Big|_0^\infty$ as

$$-x^k e^{-\lambda x}\Big|_0^\infty = \lim_{x\to\infty} -x^k e^{-\lambda x} + 0^k \cdot e^{-\lambda\cdot 0} = \lim_{x\to\infty} -x^k e^{-\lambda x} = \lim_{x\to\infty}\frac{-x^k}{e^{\lambda x}}$$

Applying L'Hopital's Rule k times yields

$$-x^k e^{-\lambda x}\Big|_0^\infty = \lim_{x\to\infty}\frac{-(k!)}{\lambda^k e^{\lambda x}} = 0$$

This implies that

$$E[X^k] = 0 + \frac{kE[X^{k-1}]}{\lambda} = \frac{k}{\lambda}E[X^{k-1}]$$

Thus, our recurrence relation is

$$E[X^k] = \begin{cases} \frac{1}{\lambda} \text{ if } k = 1 \\ \frac{kE[X^{k-1}]}{\lambda} \text{ otherwise.} \end{cases}$$

Claim: $E[X^k] = \frac{k!}{\lambda^k}$ for all $k \in \mathbb{N}$.
Proof. We induct on k.
Base Case: $k = 1$, we already found that

$$E[X^k] = E[X] = \frac{1}{\lambda} = \frac{1!}{\lambda^1} = \frac{k!}{\lambda^k}$$

so the claim holds for the base case.
Inductive Hypothesis: Assume that $E[X^k] = \frac{k!}{\lambda^k}$ for all $1 \leq k \leq n$.
Inductive Step: Consider $k = n + 1$. From our recurrence relation, we know that

$$E[X^k] = E[X^{n+1}] = \frac{(n+1)E[X^n]}{\lambda}$$

From our inductive hypothesis, we know $E[X^n] = \frac{n!}{\lambda^n}$. This directly implies that

$$E[X^k] = E[X^{n+1}] = \frac{(n+1)\frac{n!}{\lambda^n}}{\lambda} = \frac{(n+1)!}{\lambda^{n+1}} = \frac{k!}{\lambda^k}$$

which is exactly what we want to show. The conclusion that

$$E[X^k] = \frac{k!}{\lambda^k}$$

follows by induction for all $k \in \mathbb{N}$.
Thus, the higher moments of an exponential random variable X are defined by

$$\mu_k = E[X^k] = \frac{k!}{\lambda^k}$$

for all $k \in \mathbb{N}$.

8. Let $X \sim \text{Binomial}(n, p)$.

   (a) Fix $k = 1, 2, 3, \ldots$. Prove that $\lim_{n\to\infty} E[X^k]/n^k = p^k$. (Hint: recall $E[X^k] = \left(t\frac{d}{dt}\right)^k G_X(t)\Big|_{t=1}$.)

   (b) Let $Z = (X - \mu)/\sigma$. Suppose also that $p = 1/2$. Prove that $E[Z^{2k+1}] = 0$ for all $k = 1, 2, 3, \ldots$.

   (c) Show that $\lim_{n\to\infty} E[(X - E[X])^{2k}]/n^{2k} = 0$ for $k = 1, 2, 3, \ldots$. (This shows that n^{2k} is not the right "scale factor" for this random variable. We will later find that n^k *is* the right scale factor in the sense that $\lim_{n\to\infty} E[(X - E[X])^{2k})/n^k \neq 0$ is finite.)

Solution.

(a) Since $X \sim Binomial(n, p)$, we know X has probability generating function

$$G_X(t) = (pt + q)^n$$

where $q = 1 - p$. Applying the hint, we find that

$$\lim_{n\to\infty} \frac{E[X^k]}{n^k} = \lim_{n\to\infty} \frac{\left(t\frac{d}{dt}\right)^k G_X(t)\Big|_{t=1}}{n^k} = \lim_{n\to\infty} \frac{\left(t\frac{d}{dt}\right)^k (pt+q)^n\Big|_{t=1}}{n^k}$$

Note:

$$\left(t\frac{d}{dt}\right)^k G_X(t)\Big|_{t=1}$$

is a polynomial in n. To compute $\left(t\frac{d}{dt}\right)^k (pt+q)^n\Big|_{t=1}$, the product rule is applied repeatedly. One of the terms in the resulting sum results from $(pt+q)^n$ being differentiated k times then evaluated at $t = 1$. We can directly compute that

$$\begin{aligned} \left(\frac{d}{dt}\right)^k (pt+q)^n\Big|_{t=1} &= n(n-1)\cdots(n-k+1)p^k(pt+q)^{n-k}\Big|_{t=1} \\ &= n(n-1)\cdots(n-k+1)p^k = n^k p^k + f(n) \end{aligned}$$

where $f(n)$ is a polynomial in n of degree $k-1$. All other terms in the sum resulting from repeatedly applying the product rule to calculate $\left(t\frac{d}{dt}\right)^k (pt+q)^n\Big|_{t=1}$ result from differentiating $(pt+q)^n$ less than k times and differentiating at^i at least 1 time (for some $i \in [k]$ and a constant W.R.T. t and n). Differentiating at^i with respect to t does not increase the degree of n. Differentiating $(pt+q)^n$ increases the degree of n by one each time. Therefore, the degree of n is maximally $k-1$ in all such terms. Therefore, the term with the highest degree in $\left(t\frac{d}{dt}\right)^k (pt+q)^n\Big|_{t=1}$ is the term resulting from differentiating $(pt+q)^n$ k times. This term has a degree of k, so we know

$$\left(t\frac{d}{dt}\right)^k (pt+q)^n\Big|_{t=1}$$

itself is a polynomial in n with degree k for all $k \in \mathbb{N}$. Moreover, since the only term with degree k is $n^k p^k$, we know

$$\left(t\frac{d}{dt}\right)^k (pt+q)^n\Big|_{t=1} = n^k p^k + a_{k-1}n^{k-1} + \cdots + a_1 n + a_0$$

where $a_{k-1}, \cdots, a_0$ are constants W.R.T n and t. This implies that

$$\begin{aligned} \lim_{n\to\infty} \frac{E[X^k]}{n^k} &= \lim_{n\to\infty} \frac{\left(t\frac{d}{dt}\right)^k (pt+q)^n\Big|_{t=1}}{n^k} \\ &= \lim_{n\to\infty} \frac{n^k p^k + a_{k-1}n^{k-1} + \cdots + a_1 n + a_0}{n^k} \end{aligned}$$

Applying L'Hopital's Rule k times yields

$$\lim_{n\to\infty}\frac{E[X^k]}{n^k}=\lim_{n\to\infty}\frac{k!p^k}{k!}=p^k$$

for all $k\in\mathbb{N}$ since the kth derivative of $a_{k-1}n^{k-1}+\cdots+a_1n+a_0$ is 0. This concludes the proof that

$$\lim_{n\to\infty}\frac{E[X^k]}{n^k}=p^k$$

for all $k\in\mathbb{N}$.

(b) Since $p=\frac{1}{2}$, we know

$$E[X]=\mu=np=n\frac{1}{2}=\frac{n}{2}$$

and

$$SD(X)=\sqrt{Var(X)}=\sqrt{np(1-p)}=\sqrt{\frac{n}{4}}=\frac{\sqrt{n}}{2}$$

so

$$Z=\frac{X-\mu}{\sigma}=\frac{X-\frac{n}{2}}{\frac{\sqrt{n}}{2}}$$

Applying the definition of expected value and the probability mass function for X, we find

$$\begin{aligned}E[Z^k]=&\quad\sum_{i=0}^{n}\big(\frac{i-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\big)^k\mathbb{P}(Z=\frac{i-\frac{n}{2}}{\frac{\sqrt{n}}{2}})=\sum_{i=0}^{n}\big(\frac{i-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\big)^k\mathbb{P}(X=i)\\=&\quad\sum_{i=0}^{n}\big(\frac{i-\frac{n}{2}}{\frac{\sqrt{n}}{2}}\big)^k\binom{n}{i}(\frac{1}{2})^i(\frac{1}{2})^{n-i}=\frac{2^k}{2^n\sqrt{n}^k}\sum_{i=0}^{n}\big(i-\frac{n}{2}\big)^k\binom{n}{i}\end{aligned}$$

If n is odd, we can pair each term corresponding to i with the term corresponding to $n-i$ to find

$$\begin{aligned}E[Z^k]=&\quad\frac{2^k}{2^n\sqrt{n}^k}\sum_{i=0}^{\frac{n-1}{2}}\binom{n}{i}\big(i-\frac{n}{2}\big)^k+\binom{n}{n-i}\big((n-i)-\frac{n}{2}\big)^k\\=&\quad\frac{2^k}{2^n\sqrt{n}^k}\sum_{i=0}^{\frac{n-1}{2}}\binom{n}{i}\left((i-\frac{n}{2})^k+(-1)^k(i-\frac{n}{2})^k\right)\end{aligned}$$

This implies, for odd n

$$E[Z^{2k+1}]=\frac{2^{2k+1}}{2^n\sqrt{n}^{2k+1}}\sum_{i=0}^{\frac{n-1}{2}}\binom{n}{i}\left((i-\frac{n}{2})^{2k+1}-(i-\frac{n}{2})^{2k+1}\right)=\frac{2^{2k+1}}{2^n\sqrt{n}^{2k+1}}\sum_{i=0}^{\frac{n-1}{2}}\binom{n}{i}\cdot 0=0$$

since $(-1)^{2k+1}=-1$ for all $k\in\mathbb{N}$.
For even n, we can pair up the terms in a similar way, with the exception of the $i=\frac{n}{2}$ term. This yields

$$\begin{aligned}E[X^k]=&\quad\frac{2^k}{2^n\sqrt{n}^k}\left((\frac{n}{2}-\frac{n}{2})^k\binom{n}{\frac{n}{2}}+\sum_{i=0}^{\frac{n}{2}-1}\binom{n}{i}\left((i-\frac{n}{2})^k+(-1)^k(i-\frac{n}{2})^k\right)\right)\\=&\quad\frac{2^k}{2^n\sqrt{n}^k}\sum_{i=0}^{\frac{n}{2}-1}\binom{n}{i}\left((i-\frac{n}{2})^k+(-1)^k(i-\frac{n}{2})^k\right)\end{aligned}$$

This implies, for even n

$$E[Z^{2k+1}] = \frac{2^{2k+1}}{2^n\sqrt{n}^{2k+1}} \sum_{i=0}^{\frac{n}{2}-1} \binom{n}{i} \left((i-\frac{n}{2})^{2k+1} - (i-\frac{n}{2})^{2k+1} \right) = \frac{2^{2k+1}}{2^n\sqrt{n}^{2k+1}} \sum_{i=0}^{\frac{n}{2}-1} \binom{n}{i} \cdot 0 = 0$$

since $(-1)^{2k+1} = -1$ for all $k \in \mathbb{N}$.
Thus, we have shown that, for any $n \in \mathbb{N}$,

$$E[Z^{2k+1}] = 0$$

for all $k \in \mathbb{N}$, which completes the proof.

(c) We can apply the binomial theorem to find

$$E[(X - E[X])^{2k}] = E[\sum_{i=0}^{2k} \binom{n}{i} X^i (-E[X])^{2k-i}]$$

Applying linearity of expectation yields

$$E[(X - E[X])^{2k}] = \sum_{i=0}^{2k} E[\binom{n}{i} X^i (-E[X])^{2k-i}] = \sum_{i=0}^{2k} \binom{n}{i} (-1)^{2k-i} (np)^{2k-i} E[X^i]$$

From part (a), we know that

$$E[X^i] = n^i p^i + a_{i-1} n^{i-1} + \cdots a_1 n + a_0$$

which implies that

$$E[(X - E[X])^{2k}] = \sum_{i=0}^{2k} \binom{n}{i} (-1)^{2k-i} (np)^{2k-i} (n^i p^i + a_{i-1} n^{i-1} + \cdots a_1 n + a_0)$$

Each term in the sum is clearly a polynomial in n of degree $2k - i + i = 2k$, so $E[(X - E[X])^k]$ has degree of at most $2k$. Furthermore, $E[(X - E[X])^k]$ is actually a polynomial in n of degree at most $2k - 1$. We can see this by examining the coefficient of n^{2k}:

$$\begin{aligned} [n^{2k}]E[(X - E[X])^k] = &\quad [n^{2k}] \sum_{i=0}^{2k} \binom{n}{i} (-1)^{2k-i} (np)^{2k-i} (n^i p^i + a_{i-1} n^{i-1} + \cdots a_1 n + a_0) \\ = &\quad p^{2k} \sum_{i=0}^{2k} \binom{n}{i} (-1)^{2k-i} \end{aligned}$$

Note: We can multiply the each of the summands of the bottom sum by $1 = 1^i$ and apply the binomial theorem again to obtain

$$[n^{2k}]E[(X - E[X])^k] = p^{2k} \sum_{i=0}^{2k} \binom{n}{i} 1^i (-1)^{2k-i} = p^{2k}(1 + (-1))^{2k} = p^{2k} \cdot 0^{2k} = 0$$

Since the coefficient of n^{2k} is 0 in $E[(X - E[X])^{2k}]$, and we know that $E[(X - E[X])^{2k}]$ is a polynomial in n of degree at most $2k - 1$. We can then write $E[(X - E[X])^{2k}]$ as

$$E[(X - E[X])^{2k}] = a_{2k-1} n^{2k-1} + \cdots + a_1 n + a_0$$

for constants $a_0, \cdots, a_{2k-1}$ W.R.T n. This implies that

$$\lim_{n\to\infty} \frac{E[(X - E[X])^{2k}]}{n^{2k}} = \lim_{n\to\infty} \frac{a_{2k-1} n^{2k-1} + \cdots + a_1 n + a_0}{n^{2k}}$$

Applying L'Hopital's Rule $2k$ times W.R.T n yields

$$\lim_{n\to\infty} \frac{E[(X - E[X])^{2k}]}{n^{2k}} = \lim_{n\to\infty} \frac{0}{(2k)!} = 0$$

for all $k \in \mathbb{N}$ since

$$(\frac{d}{dt}^{2k}) f(n) = 0$$

for any polynomial f in n of degree at most $2k - 1$. This concludes the proof that

$$\lim_{n\to\infty} \frac{E[(X - E[X])^{2k}]}{n^{2k}} = 0$$

for all $k \in \mathbb{N}$.

Assignment 11

Math 407 (Swanson) – Spring 2023
Homework 1
Due Friday 1/13, 11:59pm

Name: Emerson Kahle Section: 39981

- You must upload your solutions to Gradescope as **one single, high-quality PDF**. You can convert paper-based work to a high-quality PDF using a scanning app for mobile devices, such as Adobe Scan (free, available for iOS and Android, can do multiple pages) or many others. If necessary, you can combine or merge multiple PDF's into a single PDF using a variety of services, such as Adobe Acrobat's cloud-based merge tool.
- After you upload, you must match each question with its corresponding page using Gradescope's interface. This allows graders to spend more time giving you feedback instead of hunting through submissions.
- Answers without supporting work will receive no credit. Show your work.
- You are encouraged to work together on homework, but **you must write up your solutions separately in your own words**. Copying from your fellow students or other sources is a serious academic integrity violation. In particular, you may not use "tutoring" services which simply provide answers.
- You are encouraged to typeset your solutions in LaTeX. Source code has been provided on Blackboard. Overleaf is a popular cloud-based editor.
- Problem numbers refer to the course textbook, though the problems may have been modified significantly.

1. (Ross, P5.7) The density function of X is given by

$$f(x) = \begin{cases} a + bx^2 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

If $E[X] = \frac{3}{5}$, find a and b.

Solution.
By the definition of the expected value of a continuous random variable,

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

We can use the definition of $f(x)$ to compute that

$$\begin{aligned} E[X] = & \int_{-\infty}^{0} xf(x)dx + \int_{0}^{1} xf(x)dx + \int_{1}^{\infty} xf(x)dx \\ = & \int_{-\infty}^{0} x \cdot 0dx + \int_{0}^{1} x(a + bx^2)dx + \int_{1}^{\infty} x \cdot 0dx \\ = & 0 + \int_{0}^{1} x(a + bx^2)dx + 0 \\ = & \int_{0}^{1} ax + bx^3 dx \\ = & \left(\frac{ax^2}{2} + \frac{bx^4}{4}\right)\Bigg|_0^1 = \left(\frac{a \cdot 1^2}{2} + \frac{b \cdot 1^4}{4}\right) - \left(\frac{a \cdot 0^2}{2} + \frac{b \cdot 0^4}{4}\right) = \frac{a}{2} + \frac{b}{4} \end{aligned}$$

By the definition of the probability density function, we know that

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

and we can compute that

$$1 = \int_{-\infty}^{\infty} f(x)dx = \int_{0}^{1} a + bx^2 dx = \left(ax + \frac{bx^3}{3}\right)\Bigg|_0^1 = \left(a \cdot 1 + \frac{b \cdot 1^3}{3}\right) - \left(a \cdot 0 + \frac{b \cdot 0}{3}\right) = a + \frac{b}{3}$$

Since we are given $E[X] = \frac{3}{5}$, we can now form a system of two linear equations in a and b:

$$\begin{cases} \frac{3}{5} = \frac{a}{2} + \frac{b}{4} \\ 1 = a + \frac{b}{3} \end{cases}$$

Solving the system yields:

$$\begin{aligned} 1 = a + \frac{b}{3} \quad & \implies \quad a = 1 - \frac{b}{3} \implies \\ \frac{3}{5} = \frac{1 - \frac{b}{3}}{2} + \frac{b}{4} = \frac{1}{2} - \frac{b}{6} + \frac{b}{4} \quad & \implies \quad \frac{3}{5} - \frac{1}{2} = \frac{b}{4} - \frac{b}{6} \implies \\ \frac{6}{10} - \frac{5}{10} = \frac{3b}{12} - \frac{2b}{12} \quad & \implies \quad \frac{1}{10} = \frac{b}{12} \implies \\ b = \frac{12}{10} = \frac{6}{5} \quad & \implies \quad a = 1 - \frac{6}{5}\frac{1}{3} = 1 - \frac{2}{5} = \frac{3}{5} \end{aligned}$$

Thus, we have found that $a = \frac{3}{5}$ and $b = \frac{6}{5}$.

2. (Ross, P5.42) If X is uniformly distributed on $(0, 1)$, find the distribution of $Y = e^X$.

We want to find a probability density function $f_Y(x)$ for the random variable Y. We can start by trying to find a cumulative distribution function $F_Y(x)$ for Y. By the definition of the cumulative distribution function, we know

$$F_Y(x) = \mathbb{P}(Y \leq x) = \mathbb{P}(e^X \leq x)$$

Since

$$e^X \leq x \iff X \leq ln(x)$$

we know

$$F_Y(x) = \mathbb{P}(X \leq ln(x)) = \int_{-\infty}^{ln(x)} f_X(x)dx$$

where $f_X(x)$ is the probability density function of X. Since $X \sim ContinuousUniform([0, 1])$, we know it has probability density function

$$f_X(x) = \frac{1}{1-0}1_{[0,1]}(x) = 1_{[0,1]}(x) = \begin{cases} 1 \text{ if } 0 \leq x \leq 1 \\ 0 \text{ otherwise.} \end{cases}$$

which allows us to compute that

$$F_Y(x) = \int_{-\infty}^{0} 0dx + \int_{0}^{ln(x)} 1_{[0,1]}(x)dx = \int_{0}^{ln(x)} 1_{[0,1]}(x)dx = \begin{cases} 0 \text{ if } x < 1 \\ \int_0^{ln(x)} 1dx = ln(x) \text{ if } 1 \leq x \leq e \\ \int_0^1 1dx + \int_1^{ln(x)} 0dx = 1 \text{ otherwise.} \end{cases}$$

Since the probability density function of Y is $f_Y(x) = F'_Y(x)$, we know that

$$f_Y(x) = F'_Y(x) = \begin{cases} \frac{1}{x} \text{ if } 1 \leq x \leq e \\ 0 \text{ otherwise.} \end{cases}$$

Thus, the distribution of $Y = e^X$ can be described by its CDF

$$F_Y(x) = \begin{cases} 0 \text{ if } x < 1 \\ ln(x) \text{ if } 1 \leq x \leq e \\ 1 \text{ otherwise.} \end{cases}$$

and its PDF

$$f_Y(x) = \begin{cases} \frac{1}{x} \text{ if } 1 \leq x \leq e \\ 0 \text{ otherwise.} \end{cases}$$

3. Recall our second interpretation of Bertrand's paradox from lecture. Specifically, L is the length of a randomly selected chord on a circle of radius r, where the chord has been chosen using i.i.d. Uniform($[0°, 360°]$) random variables θ_1, θ_2. Here θ_i is the polar angle made by one of the chord's endpoints on the circumference of the circle.

   (a) Compute and graph the CDF of L explicitly.

   (b) Compute and graph the PDF of L explicitly.

   Solution.

   (a) We want to find the CDF of L

   $$F_L(x) = \mathbb{P}(L \leq x)$$

   Let's express L in terms of a Continuous Uniform random variable. First, by symmetry, we can let $\theta_2 = 0°$, as we did in lecture. Now, we can express L in terms of θ_1. If $\theta_1 <= 180$, we can create a triangle on the top half of the circle with two sides of length r, one side of length L and an angle of θ_1 between the two sides of length r. We can split the angle pf size θ_1 in half to produce two right triangles, each with one side of length r, one side of length $\frac{L}{2}$, and one angle of size $\frac{\theta_1}{2}$ such that the side of length $\frac{L}{2}$ is opposite to the angle and the side of length r is the hypotenuse. From the definition of $sin(x)$, we know

   $$sin(\frac{\theta_1}{2}) = \frac{L}{2r} \implies L = 2rsin(\frac{\theta_1}{2})$$

   for all $0 \leq \theta_1 \leq 180$. If $\theta_1 > 180$, we can create a similar triangle on the bottom half of the circle with two sides of length r, one side of length L, and an angle of size $360 - \theta_1$ between the two sides of length r. We can now split the angle of size $360 - \theta_1$ to create two right triangles, each with one side of length r, one side of length $\frac{L}{2}$, and one angle of size $\frac{360-\theta_1}{2}$ such that the side of length $\frac{L}{2}$ is opposite to the angle and the side of length r is the hypotenuse. From the definition of $sin(x)$, we know

   $$sin(\frac{360 - \theta_1}{2}) = \frac{L}{2r} \implies L = 2rsin(\frac{360 - \theta_1}{2})$$

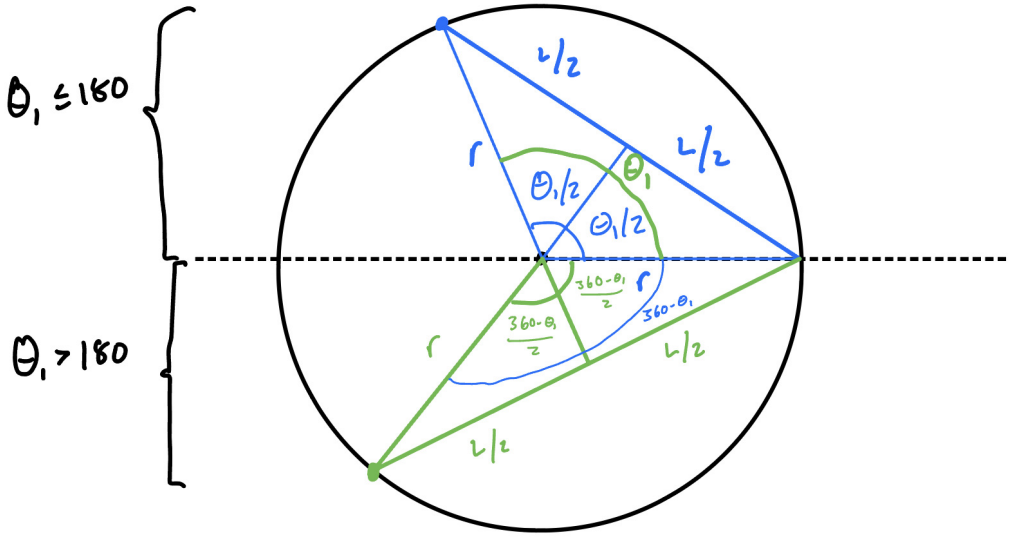
   for all $180 < \theta_1 \leq 360$. Since $360 - \theta_1$ ranges from 0 to 180 as θ_1 ranges from 180 to 360, we can define a new variable

   $$\theta = \begin{cases} \theta_1 \text{ if } 0 \leq \theta_1 \leq 180 \\ 360 - \theta_1 \text{ if } 180 < \theta_1 \leq 360 \end{cases}$$

   and, since $\theta_1 \sim ContinuousUniform([0°, 360°])$, we know $\theta \sim ContinuousUniform([0°, 180°])$, and we have

   $$L = 2rsin(\frac{\theta}{2})$$

   for all $0 \leq \theta \leq 180$. A visual demonstration of the geometric description is provided below:



Thus, we have

$$F_L(x) = \mathbb{P}(L \leq x) = \mathbb{P}(2rsin(\frac{\theta}{2}) \leq x)$$

We know

$$2rsin(\frac{\theta}{2}) \leq x \iff sin(\frac{\theta}{2}) \leq \frac{x}{2r} \iff \theta \leq 2arcsin(\frac{x}{2r})$$

which implies

$$F_L(x) = \mathbb{P}(\theta \leq 2arcsin(\frac{x}{2r}) = F_\theta(2arcsin(\frac{x}{2r}))$$

where $F_\theta(x)$ is the cumulative distribution function of θ.
Since $\theta \sim ContinuousUniform([0°, 180°])$, we know θ has PDF

$$f_\theta(x) = \begin{cases} \frac{1}{180} \text{ if } 0 \leq x \leq 180 \\ 0 \text{ otherwise} \end{cases}$$

which implies θ has CDF

$$F_\theta(x) = \begin{cases} 0 \text{ if } x < 0 \\ \int_0^x \frac{1}{180} dx = \frac{x}{180} \text{ if } 0 \leq x \leq 180 \\ 1 \text{ otherwise.} \end{cases}$$

which implies that L has CDF

$$F_L(x) = F_\theta(2arcsin(\frac{x}{2r})) = \begin{cases} 0 \text{ if } 2arcsin(\frac{x}{2r}) < 0 \\ \frac{arcsin(\frac{x}{2r})}{90} \text{ if } 0 \leq 2arcsin(\frac{x}{2r}) \leq 180 \\ 1 \text{ otherwise.} \end{cases}$$

We know that

$$2arcsin(\frac{x}{2r}) < 0 \iff \frac{x}{2r} < sin(\frac{0}{2}) = 0 \iff x < 0$$

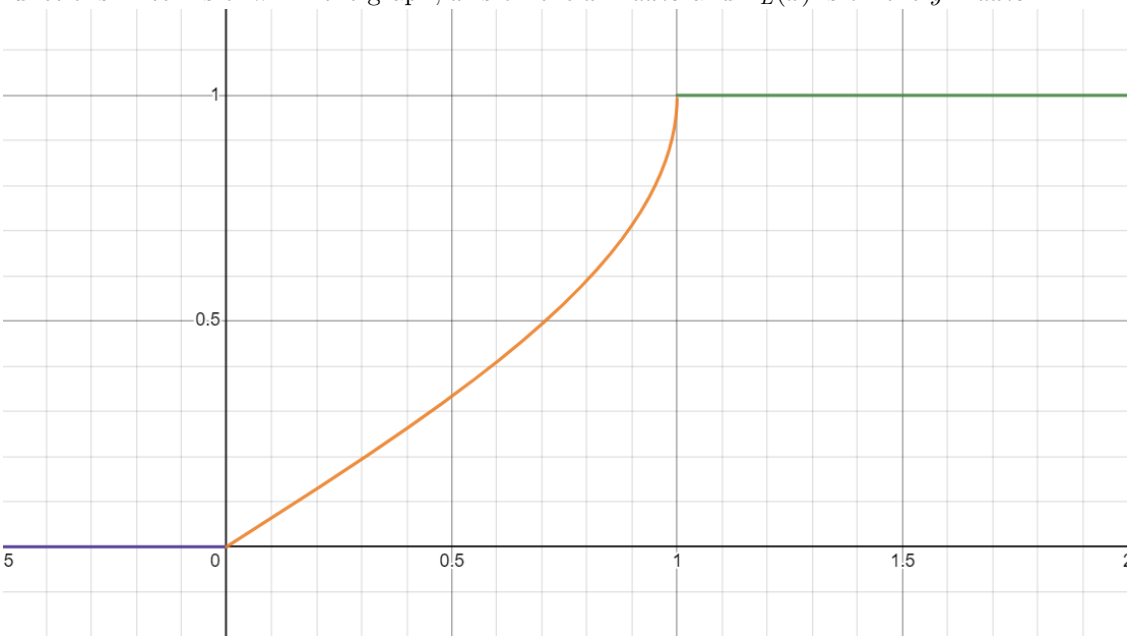
We also know that

$$0 \leq 2arcsin(\frac{x}{2r}) \leq 180 \iff 0 \leq arcsin(\frac{x}{2r}) \leq 90 \iff 0 \leq \frac{x}{2r} \leq 1 \iff 0 \leq x \leq 2r$$

So we can rewrite our CDF as

$$F_L(x) = \begin{cases} 0 \text{ if } x < 0 \\ \frac{arcsin(\frac{x}{2r})}{90} \text{ if } 0 \leq x \leq 2r \\ 1 \text{ otherwise.} \end{cases}$$

This is our explicit expression for the CDF of L. We graph $F_L(x)$ below with $r = 0.5$ for the best visual experience. We also set $180 = \pi$ because the graphing software computes trigonometric functions in terms of π. In the graph, x is on the $x - axis$ and $F_L(x)$ is on the $y - axis$:



(b) Since we already computed the CDF of L, and we know the PDF of L

$$f_L(x) = F_L'(x)$$

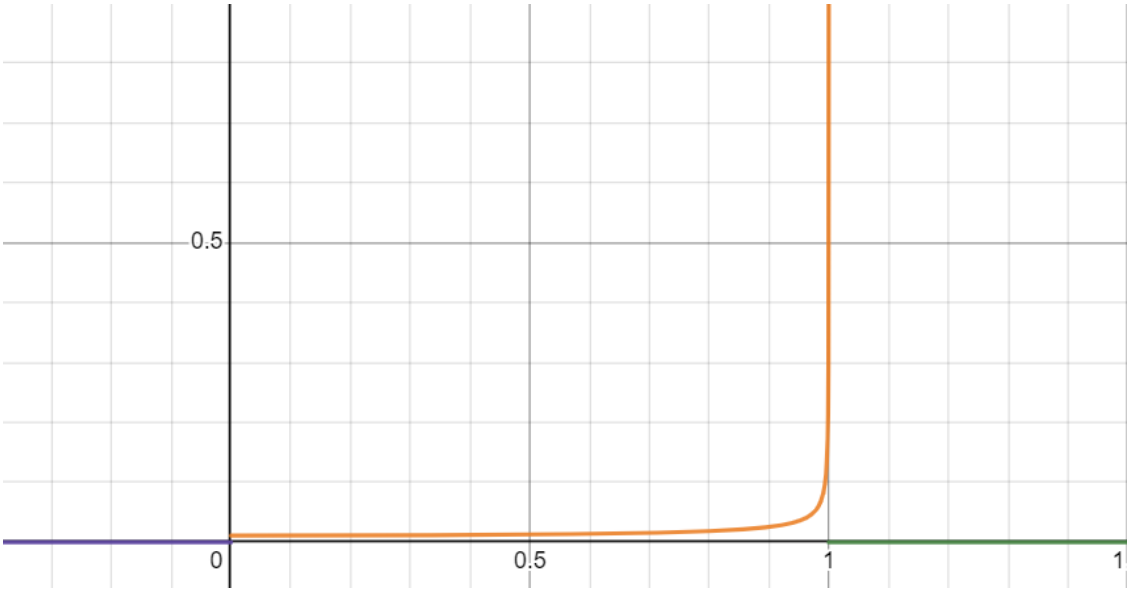
we can easily compute the PDF of L. We know that

$$\frac{d}{dx}\left(\frac{arcsin(\frac{x}{2r})}{90}\right) = \frac{1}{90}\frac{1}{\sqrt{1-(\frac{x}{2r})^2}}\frac{1}{2r} = \frac{1}{180r\sqrt{1-\frac{x^2}{4r^2}}}$$

so we can conclude that the PDF of L is

$$f_L(x) = \begin{cases} \frac{1}{180r\sqrt{1-\frac{x^2}{4r^2}}} \text{ if } 0 \leq x \leq 2r \\ 0 \text{ otherwise.} \end{cases}$$

This is our explicit expression for the PDF of L. We graph $f_L(x)$ below with $r = 0.5$ for the best visual experience. In the graph, x is on the $x - axis$ and $f_L(x)$ is on the $y - axis$:


0.5
0
0.5
1
1

4. A third interpretation of Bertrand's paradox is as follows. Uniformly choose a point P in the interior of a circle of radius r. Now interpret P as the midpoint of a chord of length L.

   (a) Let $P = (X, Y)$ where X and Y are themselves random variables. Here $-r \leq X, Y \leq r$. What is the joint probability density function of X and Y?

   (b) Let m be the edge length of an equilateral triangle inscribed in the circle. Compute $P(L > m)$ with this third interpretation.

   (c) How does your result for (b) compare to the results from lecture? Give a qualitative explanation for which of the three probabilities is highest and which is lowest.

   Solution.

   (a) By the equation of a circle, we know $P = (X, Y)$ is in the interior of a circle of radius $r \iff X^2 + Y^2 \leq r^2 \iff |Y| \leq \sqrt{r^2 - X^2}, -r \leq X \leq r, \iff -\sqrt{r^2 - X^2} \leq Y \leq \sqrt{r^2 - X^2}, -r \leq X \leq r$. Thus, we know that, for any point $P = (X, Y)$ in the interior of the circle of radius r,

$$\mathbb{P}(-r \leq X \leq r, -\sqrt{r^2 - X^2} \leq Y \leq \sqrt{r^2 - X^2}) = 1$$

   By the definition of the joint probability density function of X and Y, we know

$$1 = \mathbb{P}(-r \leq X \leq r, -\sqrt{r^2 - X^2} \leq Y \leq \sqrt{r^2 - X^2}) = \int_{-r}^{r} \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} f_{X,Y}(x, y) dy dx$$

   Since the point P is chosen uniformly at random, the probability density function should be a constant c for all $P = (X, Y)$ such that P is in the interior of a circle of radius r. Thus, for the entirety of the region being integrated over, we have $f_{X,Y}(x, y) = c \in \mathbb{R}$, so we know

$$1 = \int_{-r}^{r} \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} c dy dx = c \int_{-r}^{r} \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} dy dx$$

   Since the region being integrated over is the entire circle of radius r, we can convert to polar coordinates to find

$$\int_{-r}^{r} \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} = \int_0^{2\pi} \int_0^r r dr d\theta = \int_0^{2\pi} \frac{r^2}{2} = \frac{r^2}{2}\theta \Big|_0^{2\pi} = \frac{2\pi r^2}{2} = \pi r^2$$

   This makes sense because integrating the constant function $f = 1$ over a region equal to the entire circle of radius r should return the area of the circle, which is πr^2. We now know that

$$1 = c\pi r^2 \implies c = \frac{1}{\pi r^2}$$

   Thus, we know that X and Y have the joint probability density function $f_{X,Y}(x, y) = \frac{1}{\pi r^2}$ for all $x^2 + y^2 \leq r^2$. If $X^2 + Y^2 > r^2$, then $P = (X, Y)$ is outside of the circle, so the joint probability density function is 0. Thus, the joint probability density function of X and Y is

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{\pi r^2} \text{ if } x^2 + y^2 \leq r^2 \\ 0 \text{ otherwise.} \end{cases} = \begin{cases} \frac{1}{\pi r^2} \text{ if } -r \leq x \leq r, -\sqrt{r^2 - x^2} \leq y \leq \sqrt{r^2 - x^2} \\ 0 \text{ otherwise.} \end{cases}$$

   More concisely, we can write the joint probability density function of X and Y as

$$f_{X,Y}(x, y) = \frac{1}{\pi r^2} \cdot 1_{[-r,r]}(x) \cdot 1_{[-\sqrt{r^2-x^2}, \sqrt{r^2-x^2}]}(y)$$

(b) Suppose p_m is the midpoint of an edge of an equilateral triangle inscribed in the circle. It is clear that $L > m \iff P$ is closer to $(0,0)$ the center of the circle than p_m. Since all angles in an equilateral triangle are $60°$, we can draw a line from p_m to $(0,0)$ and from $(0,0)$ to one endpoint of the edge of the triangle corresponding to p_m to create a right triangle in which one side length is the hypotenuse r, one side length is the distance from p_m to $(0,0)$, and an angle of $30°$ sits opposite this side. If we let $d_m =$ the distance from p_m to $(0,0)$, then we know

$$sin(30°) = \frac{1}{2} = \frac{d_m}{r} \implies d_m = \frac{r}{2}$$

Thus, if we let $D =$ the distance from P to $(0,0)$, then we have $L > m \iff D < \frac{n}{2}$.
The distance from P to $(0,0)$ is

$$D = \sqrt{X^2 + Y^2}$$

so we have

$$L > m \iff \sqrt{X^2 + Y^2} < \frac{r}{2} \iff X^2 + Y^2 < (\frac{r}{2})^2$$

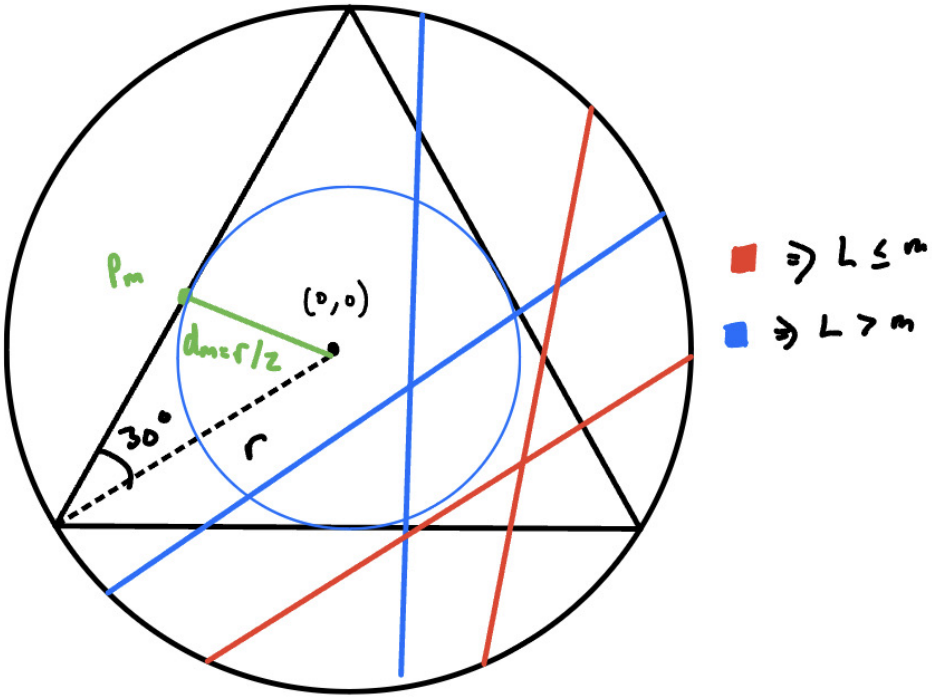
We can recognize the rightmost inequality as the set of all points inside a circle of radius $\frac{r}{2}$ centered at $(0,0)$. Call such a circle c_1, and call the circle of radius r centered at $(0,0)$ c_2. Note that the entirety of the area of c_1 is contained inside c_2. The area of c_1 is $\frac{\pi r^2}{4}$. The area of c_2 is πr^2. Since P is chosen uniformly at random, we know that

$$\mathbb{P}(L > m) = \mathbb{P}(X^2 + Y^2 < (\frac{r}{2})^2) = \frac{|X^2 + Y^2 < (\frac{r}{2})^2|}{|X^2 + Y^2 < r^2|} = \frac{Area(c_1)}{Area(c_2)} = \frac{\frac{\pi r^2}{4}}{\pi r^2} = \frac{1}{4} = 25\%$$

Thus, the probability that $L > m$ with this third interpretation is

$$\mathbb{P}(L > m) = \frac{1}{4}$$

A visual depiction of the geometric description is provided below:



(c) The result from part (b) is less than the $P(L > m) = \frac{1}{3}$ value obtained using the method of selecting endpoints randomly and less than the $\mathbb{P}(L > m) = \frac{1}{2}$ value obtained by selecting the distance from the midpoint to $(0, 0)$ randomly. Thus, $\mathbb{P}(L > m)$ is lowest when we select the midpoint of the chord randomly and $\mathbb{P}(L > m)$ is largest when we select the distance from the midpoint to $(0, 0)$ randomly. This result makes sense intuitively.

For all methods, $L > m \iff D < \frac{r}{2}$. However, the distribution of D changes significantly depending on the sampling method.

When we select the distance from the midpoint to $(0, 0)$, we have

$D \sim ContinuousUniform([0, r])$, so half of the possible chords have $D < \frac{r}{2}$. In this case, we only have $L < m$ if one variable, D is larger than $\frac{r}{2}$. This is different from the method where we select the midpoint P at random. In this case, if either $|X| > \frac{r}{2}$ or $|Y| > \frac{r}{2}$ or both, we will have $L < m$. This now accounts for $\frac{3}{4}$ of all possible chord selections, as opposed to one half. Therefore, since selecting the midpoint P randomly allows two variables to independently cause $L < m$, while selecting the distance of D to the midpoint from $(0, 0)$ only allows one variable to independently cause $L < m$, it makes sense qualitatively why $\mathbb{P}(L > m)$ is greater with the latter method. Also, when we randomly select endpoints (fixing one endpoint with $\theta_i = 0$), the length of our chord L increases from $\theta = 0°$ to $\theta = 180°$ before decreasing symmetrically. When we are halfway through the increasing portion, at $\theta = 90°$, the distance to the midpoint of our chord from $(0, 0)$ is $\frac{r}{\sqrt{2}} > \frac{r}{2}$. Therefore, since we know at least half of our chords will be too far from $(0, 0)$ for $L > m$ with the endpoint method, it makes sense that $\mathbb{P}(L > m)$ is greater for the midpoint distance D method than the endpoint method. This completes the explanation of why $\mathbb{P}(L > m)$ is greatest for the midpoint distance D method.

Since we already explained why the midpoint P method should return a lower probability $\mathbb{P}(L > m)$ than the midpoint distance D method, we just need to explain why the midpoint P method should return a lower probability $\mathbb{P}(L > m)$ than the endpoint method. We can do this by considering which cords of length $L > m$ we can choose with the two methods. If we consider some $\varepsilon > 0$, then the area of chords whose midpoint is closer than ε to $(0, 0)$ is a triangle for the endpoint method and a circle for the midpoint P method. Therefore, as $\varepsilon \to 0$, the percentage of choosable chords that satisfy $L > m$ for the midpoint P method decreases exponentially while it decreases linearly for the endpoint method. Therefore, a lower percentage of choosable chords satisfy $L > m$ for the midpoint P method than for the endpoint method. This explains why $\mathbb{P}(L > m)$ is smallest for the midpoint P method.

5. A ubiquitous probabilistic model involves a very large number of independent Bernoulli trials occuring nearly continuously, each with a very small probability of success, but where the expected number of events in a given time period is known. For instance, a large software company has many clients using its software all the time, but only rarely does any given client encounter an issue requiring them to call the company's tech support line. Suppose in a given interval of time τ, the company knows to expect λ calls on average. The company has some unknown but large number n of active users in each time interval τ and some unknown but small probability p of any active user calling in to tech support.

   (a) Show that $\lambda = np$.

   (b) Let X be the number of users interacting with the software until the first user calls in for tech support. Determine the distribution of X in terms of p.

   (c) Let Y be the amount of time until the first user calls for tech support, expressed as multiples of τ. Determine the distribution of Y in terms of p and n.

   (d) By taking the large n/small p limit, approximate $F_Y(t)$ as $n \to \infty, p \to 0, np = \lambda$.

   (e) Suppose it has been $\tau/2$ time since the last tech support call. What is the probability that the next tech support call will happen within 2τ more time?

   (f) Suppose $E \sim \text{Exponential}(\lambda)$. Show that $cE \sim \text{Exponential}(\lambda/c)$ for any $c > 0$. Give an intuitive explanation for this property.

   Solution.

   (a) Let $W =$ the number of tech support calls that the company gets in a given time interval τ. We are given that
   $$E[W] = \lambda$$
   Since the company has n active users, each of which have probability p of calling tech support, we can let $V_1, ..., V_n$ be *i.i.d. Bernoulli*(p) such that
   $$V_i = \begin{cases} 1 \text{ if the } i\text{th user calls tech support} \\ 0 \text{ otherwise.} \end{cases}$$
   Then we have
   $$W = \sum_{i=1}^{n} V_i = V_1 + \cdots + V_n$$
   so $W \sim$ *Binomial*(n, p). Thus, by the definition of the expected value of a binomial random variable, we have
   $$E[W] = np$$
   Since we are given $E[W] = \lambda$, this completes the proof that
   $$\lambda = np$$

   (b) Define $V_1, V_2, ...$ to be *i.i.d. Bernoulli*(p) such that
   $$V_i = \begin{cases} 1 \text{ if the i'th customer calls tech support} \\ 0 \text{ otherwise.} \end{cases}$$
   Then $X =$ the number of V_i's until the first customer calls tech support = the number of *i.i.d Bernoulli*(p) trials until the first success. This implies $X \sim$ *Geometric*(p). Thus, by the definition of a geometric random variable, X has probability mass function
   $$p_X(k) = \mathbb{P}(X = k) = (1-p)^{k-1}p$$

and cumulative distribution function

$$\begin{aligned} F_X(k) = \mathbb{P}(X \leq k) = & \sum_{i=1}^{\lfloor k \rfloor} (1-p)^{i-1} p = p \sum_{i=0}^{\lfloor k-1 \rfloor} (1-p)^i \\ = & \; p \frac{1-(1-p)^{\lfloor k \rfloor}}{1-(1-p)} = \begin{cases} 0 \text{ if } k < 0 \\ 1-(1-p)^{\lfloor k \rfloor} \text{ if } 0 \leq k \leq n \\ 1 \text{ otherwise.} \end{cases} \end{aligned}$$

(c) If $X = i$ for any $1 \leq i \leq n$, then the amount of time until the first user calls tech support is τ. If $X = j$ for any $n+1 \leq j \leq 2n$, then the amount of time until the first user calls tech support is 2τ. Similarly, for any $t \in \mathbb{N}$, if $X = j$ for any $(t-1)n+1 \leq j \leq tn$, then the amount of time taken until the first user calls tech support is $t\tau$. Thus, we know

$$Y = \lceil \frac{X}{n} \rceil \tau$$

so

$$\begin{aligned} \mathbb{P}(Y = t\tau) = & \; \mathbb{P}(\lceil \frac{X}{n} \rceil = t) = \mathbb{P}((t-1)n+1 \leq X \leq tn) \\ = & \sum_{i=(t-1)n+1}^{tn} \mathbb{P}(X = i) = \sum_{i=(t-1)n+1}^{tn} (1-p)^{i-1} p \\ = & \; p \sum_{i=(t-1)n}^{tn-1} (1-p)^i = p \left(\sum_{i=0}^{tn-1} (1-p)^i - \sum_{i=0}^{(t-1)n-1} (1-p)^i \right) \\ = & \; p \left(\frac{1-(1-p)^{tn}}{1-(1-p)} - \frac{1-(1-p)^{(t-1)n}}{1-(1-p)} \right) \\ = & \; p \frac{(1-p)^{(t-1)n} - (1-p)^{nt}}{p} = (1-p)^{(t-1)n} - (1-p)^{nt} \end{aligned}$$

so we know is

$$\begin{aligned} F_Y(t) = & \; \mathbb{P}(Y \leq t\tau) = \sum_{i=1}^{\lfloor t \rfloor} (1-p)^{(i-1)n} - (1-p)^{ni} \\ = & \; 1-(1-p)^n + (1-p)^n - \cdots - (1-p)^{(\lfloor t \rfloor - 1)n} + (1-p)^{(\lfloor t \rfloor - 1)n} - (1-p)^{n\lfloor t \rfloor} \\ = & \; \begin{cases} 0 \text{ if } t < 0 \\ 1-(1-p)^{n\lfloor t \rfloor} \text{ otherwise.} \end{cases} \end{aligned}$$

Thus, the probability mass function of Y is

$$p_Y(t) = \mathbb{P}(Y = t\tau) = (1-p)^{(t-1)n} - (1-p)^{nt}$$

And the cumulative distribution function of Y is

$$F_Y(t) = \begin{cases} 0 \text{ if } t < 0 \\ 1-(1-p)^{n\lfloor t \rfloor} \text{ otherwise.} \end{cases}$$

(d) We can note that $np = \lambda \implies p = \frac{\lambda}{n}$ and find

$$\lim_{\substack{n\to\infty \\ p\to 0 \\ np=\lambda}} F_Y(t) = \lim_{\substack{n\to\infty \\ p\to 0 \\ np=\lambda}} 1-(1-\frac{\lambda}{n})^{nt}$$

$$= 1 - \lim_{\substack{n\to\infty \\ p\to 0 \\ np=\lambda}} \left((1-\frac{\lambda}{n})^{n}\right)^{t}$$

$$= 1 - \left(\lim_{\substack{n\to\infty \\ p\to 0 \\ np=\lambda}} (1+(-\frac{\lambda}{n}))^{n}\right)^{t}$$

Since $e^x = \lim_{n\to\infty}(1+\frac{x}{n})^n$, we know

$$\lim_{\substack{n\to\infty \\ p\to 0 \\ np=\lambda}} F_Y(t) = 1-(e^{-\lambda})^t = 1-e^{-\lambda t}$$

Thus, by taking the large n/ small p limit, we can approximate $F_Y(t)$ as $n\to\infty$, $p\to 0$, $np=\lambda$ as

$$F_Y(t) \approx 1-e^{-\lambda t}$$

(e) We want to find

$$\mathbb{P}(Y \le \frac{5\tau}{2} | Y \ge \frac{\tau}{2})$$

We know that

$$\mathbb{P}(Y \le \frac{5\tau}{2} | Y \ge \frac{\tau}{2}) = \mathbb{P}(\frac{\tau}{2} \le Y \le \frac{5\tau}{2}) = \mathbb{P}(Y \le \frac{5\tau}{2})$$

since $\mathbb{P}(Y \ge \frac{\tau}{2}) = 1$. We can now use the CDF of Y from part (c) to find that

$$\mathbb{P}(Y \le \frac{5\tau}{2} | Y \ge \frac{\tau}{2}) = F_Y(\frac{5}{2}) = F_Y(2) = 1-(1-p)^2$$

Thus, the probability that the next tech support call will happen within 2τ more time given that it has been $\frac{\tau}{2}$ time since the last tech support call is

$$\mathbb{P}(Y \le \frac{5\tau}{2} | Y \ge \frac{\tau}{2}) = 1-(1-p)^2$$

If we want to approximate the probability for large n and small p, using the fact that the company knows to expect λ calls in any given interval of time τ, we find that

$$\mathbb{P}(Y \le \frac{5\tau}{2} | Y \ge \frac{\tau}{2}) \approx 1-e^{-2\lambda}$$

(f) Since $E \sim Exponential(\lambda)$, we know that E has cumulative distribution function

$$F_E(x) = \begin{cases} 0 \text{ if } x < 0 \\ 1 - e^{-\lambda x} \text{ otherwise.} \end{cases}$$

By the definition of the cumulative distribution function, we know cE has CDF

$$\begin{aligned} F_{cE}(x) = & \quad \mathbb{P}(cE \leq x) = \mathbb{P}(E \leq \frac{x}{c}) = F_E(\frac{x}{c}) \\ = & \quad \begin{cases} 0 \text{ if } x < 0 \\ 1 - e^{-\lambda(\frac{x}{c})} \text{ otherwise.} \end{cases} \end{aligned}$$

Therefore cE has cumulative distribution function

$$F_{cE}(x) = \begin{cases} 0 \text{ if } x < 0 \\ 1 - e^{-(\frac{\lambda}{c})x} \text{ otherwise.} \end{cases}$$

so cE has probability density function

$$f_{cE}(x) = \begin{cases} 0 \text{ if } x < 0 \\ \frac{\lambda}{c} e^{-\frac{\lambda}{c}x} \text{ otherwise.} \end{cases}$$

By the definition of an Exponential random variable, we know $cE \sim Exponential(\frac{\lambda}{c})$.

Intuitively, this property makes sense because multiplying E by a positive constant c does not change the shape of the distribution of E itself, it only stretches it by a factor of c. Therefore, if E is an exponential random variable, cE should also be an exponential random variable. In order for cE to have a valid probability distribution, the $\int_{-\infty}^{\infty} f_{cE}(x)dx = 1$ must be true. Since we scaled every value in E by a factor of c to produce cE, we must also divide the parameter λ by a factor of c to produce a valid exponential probability distribution. Thus, the result is intuitive because cE just scales all values of E by a factor of c, so cE should still be an exponential distribution, but we need to scale its parameter by a factor of $\frac{1}{c}$ to satisfy the axioms of probability.

6. (Ross, TE5.30) Let X have probability density f_X. Find the probability density function of the random variable Y defined by $Y = aX + b$.

Solution.
Since X has probability density f_X, we know

$$F_X(x) = \int_{-\infty}^{x} f_X(k)dk$$

First, we will compute the cumulative distribution function of Y, $F_Y(x)$. By the definition of the CDF, and since $Y = aX + b$, we know

$$F_Y(x) = \mathbb{P}(Y \leq X) = \mathbb{P}(aX + b \leq x)$$

If $a > 0$, we have

$$F_Y(x) = \mathbb{P}(X \leq \frac{x-b}{a}) = F_X(\frac{x-b}{a})$$

so for all $a > 0$ we have

$$F_Y(x) = F_X(\frac{x-b}{a}) = \int_{-\infty}^{\frac{x-b}{a}} f_X(k)dk$$

We can take the derivative to find that, for any $a > 0$, the probability density function of Y is

$$\frac{d}{dx}F_X(\frac{x-b}{a}) = f_X(\frac{x-b}{a})\frac{1}{a} = \frac{f_X(\frac{x-b}{a})}{a}$$

If $a < 0$, we know

$$F_Y(x) = \mathbb{P}(aX + b \leq x) = \mathbb{P}(X \geq \frac{x-b}{a}) = 1 - \mathbb{P}(X \leq \frac{x-b}{a}) = 1 - F_X(\frac{x-b}{a})$$

Once again, we can differentiate to find that, for any $a < 0$, Y has probability density function

$$f_Y(x) = \frac{d}{dx}F_Y(x) = \frac{d}{dx}(1 - F_X(\frac{x-b}{a})) = -f_X(\frac{x-b}{a})\frac{1}{a} = -\frac{f_X(\frac{x-b}{a})}{a}$$

Finally, if $a = 0$, we have $Y = aX + b = b$, so the PDF of Y is

$$f_Y(x) = 1$$

We can note that, if $a > 0$, we have

$$f_Y(x) = \frac{f_X(\frac{x-b}{a})}{a} = \frac{f_X(\frac{x-b}{a})}{|a|}$$

and if $a < 0$, we also have

$$f_Y(x) = -\frac{f_X(\frac{x-b}{a})}{a} = \frac{f_X(\frac{x-b}{a})}{|a|}$$

Thus, we can express the probability density function of $Y = aX + b$ as

$$f_Y(X) = \begin{cases} 1 \text{ if } x = 0 \\ \frac{f_X(\frac{x-b}{a})}{|a|} \text{ otherwise.} \end{cases}$$

7. (Ross, TE5.26) Let F be a continuous distribution function.

   (a) If U is uniformly distributed on $(0, 1)$, find the distribution function of $Y = F^{-1}(U)$, where F^{-1} is the inverse function of F. (That is, $y = F^{-1}(x)$ if $F(y) = x$.)

   (b) A common practical problem is to sample from a particular continuous distribution. Suppose you have a routine, `Uniform(a, b)`, which returns a uniformly random floating point number in the range (a, b). Use this routine and part (a) to construct a routine `Exponential(lambda)` which returns a floating point number sampled from the exponential distribution with rate λ.

Solution.

(a) Since $U \sim ContinuousUniform([0, 1])$, we know that U has probability density function

$$f_U(x) = \begin{cases} 1 \text{ if } 0 \leq x \leq 1 \\ 0 \text{ otherwise.} \end{cases}$$

By the definition of the cumulative distribution function, we know

$$F_Y(x) = \mathbb{P}(Y \leq X)$$

Since $Y = F^{-1}(U)$, we know that

$$F_Y(x) = \mathbb{P}(F^{-1}(U) \leq X) =$$

Since $y = F^{-1}(x) \iff F(y) = x$, we know

$$F_Y(x) = \mathbb{P}(F(F^{-1}(U)) \leq F(x)) = \mathbb{P}(U \leq F(x)) = F_U(F(x))$$

By the definition of the cumulative distribution function, we know

$$F_Y(x) = F_U(F(x)) = \int_{-\infty}^{F(x)} f_U(k)dk = \int_0^{F(x)} f_U(k)dk = \begin{cases} 0 \text{ if } F(x) < 0 \\ F(x) \text{ if } 0 \leq F(x) \leq 1 \\ 1 \text{ otherwise.} \end{cases}$$

Thus, if $Y = F^{-1}(U)$, where $U \sim ContinuousUniform([0, 1])$, then the cumulative distribution function of Y is

$$F_Y(x) = \begin{cases} 0 \text{ if } F(x) < 0 \\ F(x) \text{ if } 0 \leq F(x) \leq 1 \\ 1 \text{ otherwise.} \end{cases}$$

(b) Suppose $U \sim ContinuousUniform([0, 1])$. We need to find F^{-1} such that setting $Y = F^{-1}(U)$ results in

$$F_Y(x) = \begin{cases} 0 \text{ if } x < 0 \\ 1 - e^{-\lambda x} \text{ otherwise} \end{cases}$$

We know $Y = F^{-1}(U)$ has cumulative distribution function

$$F_Y(x) = \begin{cases} 0 \text{ if } F(x) < 0 \\ F(x) \text{ if } 0 \leq F(x) \leq 1 \\ 1 \text{ otherwise.} \end{cases}$$

so we just need to find the inverse of $F(x) = 1 - e^{-\lambda x}$. **Note:**

$$\begin{array}{lll} x \mapsto 1 - e^{-\lambda x} & \implies & x - 1 \mapsto -e^{-\lambda x} \implies \\ (1 - x) \mapsto e^{-\lambda x} & \implies & ln(1 - x) \mapsto -\lambda x \implies \\ \dfrac{-ln(1 - x)}{\lambda} \mapsto x & & \end{array}$$

so we know $F^{-1}(x) = \frac{-ln(1-x)}{\lambda}$.
Thus, we can let $Y = F^{-1}(U) = \frac{-ln(1-U)}{\lambda}$ and we know Y has cumulative distribution function

$$F_Y(x) = \begin{cases} 0 \text{ if } 1 - e^{-\lambda x} < 0 \\ 1 - e^{-\lambda x} \text{ if } 0 \leq x \end{cases}$$

so $Y \sim Exponential(\lambda)$.

To summarize, our routine to return a floating point number sampled from $Exponential(\lambda)$ is
(i) Return $\frac{-ln(1-Uniform(0,1)}{\lambda}$
Explicitly, our routine to return a floating point number sampled from $Exponential(\lambda)$ is

$$Exponential(\lambda) = \frac{-ln(1 - Uniform(0,1))}{\lambda}$$

Assignment 12

Math 407 (Swanson) – Spring 2023
Homework 1
Due Friday 1/13, 11:59pm

Name: Emerson Kahle Section: 39981

- You must upload your solutions to Gradescope as **one single, high-quality PDF**. You can convert paper-based work to a high-quality PDF using a scanning app for mobile devices, such as Adobe Scan (free, available for iOS and Android, can do multiple pages) or many others. If necessary, you can combine or merge multiple PDF's into a single PDF using a variety of services, such as Adobe Acrobat's cloud-based merge tool.
- After you upload, you must match each question with its corresponding page using Gradescope's interface. This allows graders to spend more time giving you feedback instead of hunting through submissions.
- Answers without supporting work will receive no credit. Show your work.
- You are encouraged to work together on homework, but **you must write up your solutions separately in your own words**. Copying from your fellow students or other sources is a serious academic integrity violation. In particular, you may not use "tutoring" services which simply provide answers.
- You are encouraged to typeset your solutions in LaTeX. Source code has been provided on Blackboard. Overleaf is a popular cloud-based editor.
- Problem numbers refer to the course textbook, though the problems may have been modified significantly.

1. (Ross, P6.9) The joint probability density function of X and Y is given by

$$f()x,y) = \frac{6}{7}\left(x^2 + \frac{xy}{2}\right), \qquad 0 < x < 1, 0 < y < 2.$$

(a) Verify that this is indeed a joint density function.
(b) Compute the density function of X.
(c) Find $P(X > Y)$
(d) Find $P(Y > \frac{1}{2} \mid X < \frac{1}{2})$.
(e) Find $E[Y]$.

Solution.

(a) For $f(x,y)$ to be a valid joint density function, we need

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = 1$$

Since we are given $f(x,y)$, we can compute that

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = & \quad \int_0^1\int_0^2 \frac{6}{7}\left(x^2 + \frac{xy}{2}\right)dydx = \int_0^1 \frac{6}{7}\left(x^2(2) + \frac{x(2)^2}{4}\right)dx \\ = & \quad \frac{6}{7}\int_0^1 2x^2 + xdx = \frac{6}{7}\left(\frac{2(1)^3}{3} + \frac{1^2}{2}\right) = \frac{6}{7}(\frac{4}{6} + \frac{3}{6}) = \frac{6}{7}\frac{7}{6} = 1\end{aligned}$$

By showing that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = \int_0^1\int_0^2 \frac{6}{7}\left(x^2 + \frac{xy}{2}\right)dydx = 1$$

we have completed the verification that $f(x,y)$ is indeed a joint density function.

(b) We know that the density function of X is equal to the X-marginal which is

$$\int_{-\infty}^{\infty} f(x,y)dy$$

We can directly compute that

$$\begin{aligned}\int_{-\infty}^{\infty} f(x,y)dy = & \quad \int_0^2 \frac{6}{7}\left(x^2 + \frac{xy}{2}\right)dy = \frac{6}{7}\left(x^2(2) + \frac{x(2)^2}{4}\right) \\ = & \quad \frac{6}{7}\left(2x^2 + x\right) = \frac{12x^2 + 6x}{7}\end{aligned}$$

Thus, X has the density function

$$f_X(x) = \begin{cases} \frac{12x^2+6x}{7} \text{ if } 0 < x < 1 \\ 0 \text{ otherwise.} \end{cases}$$

(c) To calculate $\mathbb{P}(X > Y)$, we only have to consider values of Y that are strictly less than X, for each value of X. Therefore, while we still integrate over all values of X, we can restrict our integration over Y to only range from $-\infty$ to X for each value of X. This allows us to compute

$$\begin{aligned}\mathbb{P}(X > Y) = & \quad \int_{-\infty}^{\infty}\int_{-\infty}^{x} \frac{6}{7}\left(x^2 + \frac{xy}{2}\right)dydx = \int_0^1\int_0^x \frac{6}{7}\left(x^2 + \frac{xy}{2}\right)dydx \\ = & \quad \frac{6}{7}\int_0^1 x^2(x) + \frac{x(x^2)}{4}dx = \frac{6}{7}\int_0^1 \frac{5x^3}{4}dx \\ = & \quad \frac{15}{14}\frac{1^4}{4} = \frac{15}{56} \approx 26.79\%\end{aligned}$$

(d) By the definition of conditional probability, we know

$$\mathbb{P}(Y > \frac{1}{2} | X < \frac{1}{2}) = \frac{\mathbb{P}(Y > \frac{1}{2}, X < \frac{1}{2})}{\mathbb{P}(X < \frac{1}{2})} \qquad (1)$$

We can use the PDF of X to compute that

$$\begin{aligned} \mathbb{P}(X < \frac{1}{2}) = & \int_{-\infty}^{\frac{1}{2}} \frac{6}{7}(2x^2 + x)dx = \frac{6}{7}\int_0^{\frac{1}{2}} 2x^2 + xdx = \frac{6}{7}\left(\frac{2(\frac{1}{2})^3}{3} + \frac{\left(\frac{1}{2}\right)^2}{2}\right) \\ = & \frac{6}{7}\left(\frac{1}{12} + \frac{1}{8}\right) = \frac{6}{7}\left(\frac{4}{48} + \frac{6}{48}\right) = \frac{6 \cdot 10}{7 \cdot 48} = \frac{10}{56} = \frac{5}{28} \end{aligned}$$

We can use the joint density function of X and Y to compute that

$$\begin{aligned} \mathbb{P}(Y > \frac{1}{2}, X < \frac{1}{2}) = & \int_0^{\frac{1}{2}} \int_{\frac{1}{2}}^2 \frac{6}{7}\left(x^2 + \frac{xy}{2}\right) dydx = \frac{6}{7}\int_0^{\frac{1}{2}} \left(2x^2 + x - (\frac{x^2}{2} + \frac{x}{16})\right) dx \\ = & \frac{6}{7}\int_0^{\frac{1}{2}} \frac{3x^2}{2} + \frac{15x}{16}dx = \frac{6}{7}\left(\frac{\frac{1}{2}^3}{2} + \frac{15(\frac{1}{2})^2}{32}\right) = \frac{6}{7}\left(\frac{1}{16} + \frac{15}{128}\right) \\ = & \frac{6}{7}\frac{23}{128} = \frac{3 \cdot 23}{7 \cdot 64} = \frac{69}{448} \end{aligned}$$

Plugging the computed values for $\mathbb{P}(X < \frac{1}{2})$ and $\mathbb{P}(Y > \frac{1}{2}, X < \frac{1}{2})$ into (1), we find

$$\mathbb{P}(Y > \frac{1}{2} | X < \frac{1}{2}) = \frac{69}{448}\frac{28}{5} = \frac{69}{80} = 86.25\%$$

(e) By the definition of expected value, we know

$$E[Y] = \int_{-\infty}^{\infty} y f_Y(y)dy \qquad (2)$$

We know that, for $0 < y < 2$

$$\begin{aligned} f_Y(y) = & \int_{-\infty}^{\infty} f(x, y)dx = \int_0^1 \frac{6}{7}\left(x^2 + \frac{xy}{2}\right) dx \\ = & \frac{6}{7}\left(\frac{1^3}{3} + \frac{1^2 y}{4}\right) = \frac{6}{7}\left(\frac{1}{3} + \frac{y}{4}\right) = \frac{2}{7} + \frac{3y}{14} \end{aligned}$$

so we have

$$f_Y(y) = \begin{cases} \frac{2}{7} + \frac{3y}{14} \text{ if } 0 < y < 2 \\ 0 \text{ otherwise.} \end{cases}$$

Plugging this into (2), we find

$$\begin{aligned} E[Y] = & \int_{-\infty}^{\infty} y f_Y(y)dy = \int_0^2 y\left(\frac{2}{7} + \frac{3y}{14}\right) dy \\ = & \int_0^2 \frac{2y}{7} + \frac{3y^2}{14}dy = \frac{4}{7} + \frac{8}{14} = \frac{4}{7} + \frac{4}{7} = \frac{8}{7} \end{aligned}$$

Thus, the expected value of Y is $\frac{8}{7}$.

2. (Ross, ST6.5) Suppose that X, Y, and Z are independent random variables that are each equally likely to be either 1 or 2. Find the probability mass function of

   (a) XYZ
   (b) $X^2 + YZ$

   Solution.

   (a) We are given that X, Y, and Z are *i.i.d* $DiscreteUniform(\{1,2\})$. Thus, they all have the same probability mass function

$$p_X(x) = p_Y(x) = p_Z(x) = \begin{cases} \frac{1}{2} \text{ if } x \in \{1,2\} \\ 0 \text{ otherwise.} \end{cases}$$

   Since X, Y, and Z can only take on values in $\{1,2\}$, we know that XYZ will be some 2^k for some $0 \leq k \leq 3$. Since X, Y, and Z are independent, we can easily compute $\mathbb{P}(XYZ = 2^k)$ for each such k. Note that there are 8 equally likely combinations of values for the variables X, Y, Z.

   (i) For $XYZ = 2^0 = 1$, we need 0 random variables to be 2, which can only happen in 1 way (all random variables are 1). Thus,

$$\begin{aligned} \mathbb{P}(XYZ = 2^0 = 1) = & \quad \mathbb{P}(X = 1, Y = 1, Z = 1) = \mathbb{P}(X = 1)\mathbb{P}(Y = 1)\mathbb{P}(Z = 1) \\ = & \quad \frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8} = 12.5\% \end{aligned}$$

   (ii) For $XYZ = 2^1 = 2$, we need exactly 1 random variable to be 2, which can happen in $\binom{3}{1}$ ways since we have three *i.i.d* $DiscreteUniform(\{1,2\})$ random variables. Thus,

$$\begin{aligned} \mathbb{P}(XYZ = 2^1 = 2) = & \quad \mathbb{P}((X = 1, Y = 1, Z = 2) \cup (X = 1, Y = 2, Z = 1) \\ & \quad \cup (X = 2, Y = 1, Z = 1)) \\ = & \quad \mathbb{P}(X = 1, Y = 1, Z = 2) + \mathbb{P}(X = 1, Y = 2, Z = 1) + \\ = & \quad \mathbb{P}(X = 2, Y = 1, Z = 1) \\ = & \quad \mathbb{P}(X = 1)\mathbb{P}(Y = 1)\mathbb{P}(Z = 2) + \mathbb{P}(X = 1)\mathbb{P}(Y = 2)\mathbb{P}(Z = 1) + \\ = & \quad \mathbb{P}(X = 2)\mathbb{P}(Y = 1)\mathbb{P}(Z = 1) \\ = & \quad \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 = \binom{3}{1}\frac{1}{8} = \frac{3}{8} = 37.5\% \end{aligned}$$

   (iii) Similarly, for $XYZ = 2^2 = 4$, we need exactly 2 random variables to be 2, which can happen in $\binom{3}{2}$ ways. Thus,

$$\begin{aligned} \mathbb{P}(XYZ = 2^2 = 4) = & \quad \mathbb{P}((X = 1, Y = 2, Z = 2) \cup (X = 2, Y = 2, Z = 1) \\ & \quad \cup (X = 2, Y = 1, Z = 2)) \\ = & \quad \mathbb{P}(X = 1, Y = 2, Z = 2) + \mathbb{P}(X = 2, Y = 2, Z = 1) + \\ = & \quad \mathbb{P}(X = 2, Y = 1, Z = 2) \\ = & \quad \mathbb{P}(X = 1)\mathbb{P}(Y = 2)\mathbb{P}(Z = 2) + \mathbb{P}(X = 2)\mathbb{P}(Y = 2)\mathbb{P}(Z = 1) + \\ = & \quad \mathbb{P}(X = 2)\mathbb{P}(Y = 1)\mathbb{P}(Z = 2) \\ = & \quad \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 = \binom{3}{2}\frac{1}{8} = \frac{3}{8} = 37.5\% \end{aligned}$$

   (iv) For $XYZ = 2^3 = 8$, we need all 3 random variables to be 2, which can only happen in 1 way. Thus,

$$\begin{aligned} \mathbb{P}(XYZ = 2^3 = 8) = & \quad \mathbb{P}(X = 2, Y = 2, Z = 2) \\ = & \quad \mathbb{P}(X = 2)\mathbb{P}(Y = 2)\mathbb{P}(Z = 2) = \frac{1}{2}^3 = \frac{1}{8} = 12.5\% \end{aligned}$$

Thus, the probability mass function of XYZ is

$$p_{XYZ}(xyz) = \begin{cases} \frac{1}{8} \text{ if } xyz \in \{1, 8\} \\ \frac{3}{8} \text{ if } xyz \in \{2, 4\} \\ 0 \text{ otherwise.} \end{cases}$$

(b) Since X can only take on values in $\{1, 2\}$, X^2 can only take on values in $\{1^2, 2^2\} = \{1, 4\}$. Also, $X^2 = 1 \iff X = 1$ and $X^2 = 4 \iff X = 2$, so we know X^2 has probability mass function

$$p_{X^2}(x^2) = \begin{cases} \frac{1}{2} \text{ if } x \in \{1, 2\} \iff x^2 \in \{1, 4\}, x > 0 \\ 0 \text{ otherwise.} \end{cases}$$

Since Y and Z can only take on values in {1,2}, we know that YZ will be some 2^k for $k \in \{0, 1, 2\}$. We can easily compute the probability that $YZ = 2^k$ for each such k:

(i) For $YZ = 1$, we need both Y and Z to equal 1, which can only happen in 1 way. Thus,

$$\mathbb{P}(YZ = 1) = \mathbb{P}(Y = 1, Z = 1) = \mathbb{P}(Y = 1)\mathbb{P}(Z = 1) = (\frac{1}{2})^2 = \frac{1}{4}$$

(ii) For $YZ = 2$, we need exactly 1 of Y and Z to be 1, which can be done in 2 ways. Thus,

$$\begin{aligned} \mathbb{P}(YZ = 2) = & \quad \mathbb{P}((Y = 2, Z = 1) \cup (Y = 1, Z = 2)) = \mathbb{P}(Y = 2, Z = 1) + \mathbb{P}(Y = 1, Z = 2) \\ = & \quad \frac{1}{4} + \frac{1}{4} = \frac{1}{2} \end{aligned}$$

(iii) For $YZ = 4$, we need neither of Y and Z to be 1, which can be done in exactly 1 way. Thus,

$$\mathbb{P}(YZ = 4) = \mathbb{P}(Y = 2, Z = 2) = \mathbb{P}(Y = 2)\mathbb{P}(Z = 2) = \frac{1}{4}$$

so YZ has probability mass function

$$P_{YZ}(yz) = \begin{cases} \frac{1}{4} \text{ if } yz \in \{1, 4\} \\ \frac{1}{2} \text{ if } yz = 2 \\ 0 \text{ otherwise.} \end{cases}$$

Since X^2 takes on values in $\{1, 4\}$ and YZ takes on values in {1, 2, 4}, $X^2 + YZ$ takes on values in $\{2, 3, 5, 6, 8\}$. Since X^2 and YZ are independent, we can easily compute $\mathbb{P}((X^2 + YZ) = a)$ for all $a \in \{2, 3, 5, 6, 8\}$:

(i) For $X^2 + YZ = 2$, we need both $X^2 = 1$ and $YZ = 1$, so we have

$$\mathbb{P}(X^2 + YZ = 2) = \mathbb{P}(X^2 = 1, YZ = 1) = \mathbb{P}(X^2 = 1)\mathbb{P}(YZ = 1) = \frac{1}{2}\frac{1}{4} = \frac{1}{8}$$

(ii) For $X^2 + YZ = 3$, we need $X^2 = 1, YZ = 2$, so we have

$$\mathbb{P}(X^2 + YZ = 3) = \mathbb{P}(X^2 = 1, YZ = 2) = \mathbb{P}(X^2 = 1)\mathbb{P}(YZ = 2) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$$

(iii) For $X^2 + YZ = 5$, we need $X^2 = 1, YZ = 4$ or $X^2 = 4, YZ = 1$. Thus, we have

$$\begin{aligned} \mathbb{P}(X^2 + YZ = 5) = & \quad \mathbb{P}((X^2 = 1, YZ = 4) \cup (X^2 = 4, YZ = 1)) \\ = & \quad \mathbb{P}(X^2 = 1, YZ = 4) + \mathbb{P}(X^2 = 4, YZ = 1) \\ = & \quad \mathbb{P}(X^2 = 1)\mathbb{P}(YZ = 4) + \mathbb{P}(X^2 = 4)\mathbb{P}(YZ = 1) \\ = & \quad \frac{1}{2}\frac{1}{4} + \frac{1}{2}\frac{1}{4} = \frac{2}{8} = \frac{1}{4} \end{aligned}$$

(iv) For $X^2 + YZ = 6$, we need $X^2 = 4$, $YZ = 2$. Thus, we have

$$\mathbb{P}(X^2 + YZ = 6) = \mathbb{P}(X^2 = 4, YZ = 2) = \mathbb{P}(X^2 = 4)\mathbb{P}(YZ = 2) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$$

(v) For $X^2 + YZ = 8$, we need $X^2 = 4, YZ = 4$. Thus, we have

$$\mathbb{P}(X^2 + YZ = 8) = \mathbb{P}(X^2 = 4, YZ = 4) = \mathbb{P}(X^2 = 4)\mathbb{P}(YZ = 4) = \frac{1}{2}\frac{1}{4} = \frac{1}{8}$$

Thus, the probability mass function of $X^2 + YZ$ is

$$p_{X^2+YZ}(x^2 + yz) = \begin{cases} \frac{1}{8} \text{ if } x^2 + yz \in \{2, 8\}, x, y, z \geq 0 \\ \frac{1}{4} \text{ if } x^2 + yz \in \{3, 5, 6\}, x, y, z \geq 0 \\ 0 \text{ otherwise.} \end{cases}$$

3. Let

$$f(x,y) = C(\sin(x+y) - \sin(x-y))1_{[0,\pi/2]}(x)1_{[0,\pi]}(y)$$

be a joint PDF. (Here 1_A is the indicator function on the set A.)

(a) Compute C.

(b) Show that the marginals X and Y are independent.

Solution.

(a) For $f(x,y)$ to be a joint PDF, we must have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = 1$$

First we can simplify $f(x,y)$. Note that, since $sin(x)$ is an odd function, $-sin(x-y) = sin(y-x)$, so we know

$$f(x,y) = C(\sin(y+x) + \sin(y-x))1_{[0,\pi/2]}(x)1_{[0,\pi]}(y)$$

Now, we can apply the sum-to-product formula for $sin(y+x) + sin(y-x)$ to find

$$\begin{aligned} f(x,y) = \quad & C(2sin(\frac{y+x+y-x}{2})cos(\frac{y+x-y+x}{2}))1_{[0,\pi/2]}(x)1_{[0,\pi]}(y) \\ = \quad & 2Csin(\frac{2y}{2})cos(\frac{2x}{2})1_{[0,\pi/2]}(x)1_{[0,\pi]}(y) = 2Csin(y)cos(x)1_{[0,\pi/2]}(x)1_{[0,\pi]}(y) \end{aligned}$$

We can plug in this simplified $f(x,y)$ to directly compute that

$$\begin{aligned} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = & \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} 2Csin(y)cos(x)1_{[0,\pi/2]}(x)1_{[0,\pi]}(y)dydx \\ = & \int_0^{\frac{\pi}{2}}\int_0^{\pi} 2Csin(y)cos(x)dydx \\ = & 2C\int_0^{\frac{\pi}{2}}\int_0^{\pi} sin(y)cos(x)dydx \\ = & 2C\int_0^{\frac{\pi}{2}} \left(-cos(\pi)cos(x) - (-cos(0)cos(x))\right)dx \\ = & 2C\int_0^{\frac{\pi}{2}} 2cos(x)dx \\ = & 4C(sin(\frac{\pi}{2}) - sin(0)) \\ = & 4C \cdot 1 = 4C \end{aligned}$$

Thus, since $f(x,y)$ is a valid joint PDF, we know

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = 4C$$

which implies

$$C = \frac{1}{4}$$

and

$$\begin{aligned} f(x,y) = & \frac{1}{4}(\sin(x+y) - \sin(x-y))1_{[0,\pi/2]}(x)1_{[0,\pi]}(y) \\ = & \frac{1}{2}sin(y)cos(x)1_{[0,\pi/2]}(x)1_{[0,\pi]}(y) \end{aligned}$$

(b) The X marginal is

$$\begin{aligned} p_X(x) = & \int_{-\infty}^{\infty} f(x,y)dy = \int_0^{\pi} \frac{1}{2} sin(y)cos(x)dy \\ = & -\frac{1}{2}cos(\pi)cos(x) - (-\frac{1}{2}cos(0)cos(x)) = \frac{cos(x)+cos(x)}{2} = \frac{2cos(x)}{2} = cos(x) \end{aligned}$$

so we have

$$p_X(x) = \begin{cases} cos(x) \text{ if } 0 < x < \frac{\pi}{2} \\ 0 \text{ otherwise.} \end{cases}$$

assuming $0 < Y < \pi$.
The Y marginal is

$$\begin{aligned} p_Y(y) = & \int_{-\infty}^{\infty} f(x,y)dx = \int_0^{\frac{\pi}{2}} \frac{1}{2} sin(y)cos(x)dx \\ = & \frac{1}{2}sin(y)sin(\frac{\pi}{2}) - \frac{1}{2}sin(y)sin(0) = \frac{sin(y)}{2} \end{aligned}$$

so we have

$$p_Y(y) = \begin{cases} \frac{sin(y)}{2} \text{ if } 0 < y < \pi \\ 0 \text{ otherwise.} \end{cases}$$

assuming $0 < X < \frac{\pi}{2}$. Multiplying $p_Y(y)$ and $p_X(x)$ together, we find

$$\begin{aligned} p_X(x)p_Y(y) = & \begin{cases} cos(x)\frac{sin(y)}{2} = \frac{1}{2}sin(y)cos(x) \text{ if } 0 < x < \frac{\pi}{2}, 0 < y < \pi \\ 0 \text{ otherwise} \end{cases} \\ = & \frac{1}{2}sin(y)cos(x)1_{[0,\frac{\pi}{2}]}(x)1_{[0,\pi]}(y) = f(x,y) \end{aligned}$$

This holds for all $-\infty < x, y < \infty$, which completes the proof that the marginals of X and Y are independent. We can also note that

$$p_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{\frac{1}{2}sin(y)cos(x)}{\frac{1}{2}sin(y)}1_{[0,\frac{\pi}{2}]}(x)1_{[0,\pi]}(y) = cos(x)1_{[0,\frac{\pi}{2}]}(x)1_{[0,\pi]}(y) = p_X(x)$$

and

$$P_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{\frac{1}{2}sin(y)cos(x)}{cos(x)}1_{[0,\frac{\pi}{2}]}(x)1_{[0,\pi]}(y) = \frac{1}{2}sin(y)1_{[0,\frac{\pi}{2}]}(x)1_{[0,\pi]}(y) = p_Y(y)$$

to conclude that the X and Y marginals are independent.

4. (Ross, P6.43) Two dice are rolled. Let X and Y denote, respectively, the largest and smallest values obtained. Compute the conditional mass function of Y given $X = i$, for $i = 1, 2, \ldots, 6$. Are X and Y independent? Why?

Solution.
We can do this directly for each i. Suppose v_i = the value on the ith roll.

(i) X = i = 1. If $v_i > 1$ for any $i \in \{1, 2\}$, then $X \neq 1$, as the value of the largest roll could not be 1. Thus, $X = 1 \implies v_i = 1$ for all $i \in \{1, 2\}$. Since both rolls must be 1, the smallest roll must be 1, so we know

$$p_{Y|X}(y|1) = \begin{cases} 1 \text{ if } y = 1 \\ 0 \text{ otherwise.} \end{cases}$$

(ii) $X = i = 2$. Since the largest roll is 2, the smallest roll could either be 2 or 1. There are 3 equally likely ways to have a largest roll of 2:

$$(v_1, v_2) \in \{(1, 2), (2, 1), (2, 2)\}$$

Of these 3 ways, $Y = 1$ in the first 2, and $Y = 2$ in the last 1. Thus, we have

$$p_{Y|X}(y|2) = \begin{cases} \frac{2}{3} \text{ if } y = 1 \\ \frac{1}{3} \text{ if } y = 2 \\ 0 \text{ otherwise.} \end{cases}$$

(iii) $X = i = 3$. Since the largest roll is 3, the smallest roll could be 1, 2, or 3. There are 5 equally likely ways to have a largest roll of 3:

$$(v_1, v_2) \in \{(1, 3), (3, 1), (2, 3), (3, 2), (3, 3)\}$$

Of these 5 ways, $Y = 1$ in 2 of them, $Y = 2$ in 2 of them, and $Y = 3$ in 1 of them. Thus, we have

$$p_{Y|X}(y|3) = \begin{cases} \frac{2}{5} \text{ if } y \in \{1, 2\} \\ \frac{1}{5} \text{ if } y = 3 \\ 0 \text{ otherwise.} \end{cases}$$

(iv) $X = i = 4$. Since the largest roll is 4, the smallest roll could be 1, 2, 3, or 4. There are 7 equally likely ways to have a largest roll of 4:

$$(v_1, v_2) \in \{(1, 4), (4, 1), (2, 4), (4, 2), (3, 4), (4, 3), (4, 4)\}$$

Of these 7 ways, $Y = 1$ in 2 of them, $Y = 2$ in 2 of them, $Y = 3$ in 2 of them, and $Y = 4$ in 1 of them. Thus, we have

$$p_{Y|X}(y|4) = \begin{cases} \frac{2}{7} \text{ if } y \in \{1, 2, 3\} \\ \frac{1}{7} \text{ if } y = 4 \\ 0 \text{ otherwise.} \end{cases}$$

(v) $X = i = 5$. Since the largest roll is 5, the smallest roll could be 1, 2, 3, 4, or 5. There are 9 equally likely ways to have a largest roll of 5:

$$(v_1, v_2) \in \{(1, 5), (5, 1), (2, 5), (5, 2), (3, 5), (5, 3), (4, 5), (5, 4), (5, 5)\}$$

Of these 9 ways $Y = 1$ in 2 of them, $Y = 2$ in 2 of them, $Y = 3$ in 2 of them, $Y = 4$ in 2 of them, and $Y = 5$ in 1 of them. Thus, we have

$$p_{Y|X}(y|5) = \begin{cases} \frac{2}{9} \text{ if } y \in \{1, 2, 3, 4\} \\ \frac{1}{9} \text{ if } y = 5 \\ 0 \text{ otherwise.} \end{cases}$$

(vi) $X = i = 6$. Since the largest roll is 6, the smallest roll could be 1, 2, 3, 4, 5, or 6. There are 11 equally likely ways to have a largest roll of 6:

$$(v_1, v_2) \in \{(1,6),(6,1),(2,6),(6,2),(3,6),(6,3),(4,6),(6,4),(5,6),(6,5),(6,6)$$

Of these 11 ways, $Y = 1$ in 2 of them, $Y = 2$ in 2 of them, $Y = 3$ in 2 of them, $Y = 4$ in 2 of them, $Y = 5$ in 2 of them, and $Y = 6$ in 1 of them. Thus, we have

$$p_{Y|X}(y|6) = \begin{cases} \frac{2}{11} \text{ if } y \in \{1,2,3,4,5\} \\ \frac{1}{11} \text{ if } y = 6 \\ 0 \text{ otherwise.} \end{cases}$$

We can also use these solutions for the individual conditional mass functions to find a general conditional mass function for $Y|X$. By the definition of conditional probability, we have

$$p_{Y|X}(y|x) = \mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)} \qquad (1)$$

If $y > x$, then $\mathbb{P}(Y = y, X = x) = 0$.
If $1 \leq y = x \leq 6$, then we must have $v_1 = v_2 = y = x$, which can only happen in 1 way. Since there are 36 equally likely sequences of dice rolls, this means $\mathbb{P}(Y = y, X = x) = \frac{1}{36}$ for all $y = x$.
If $1 \leq y < x \leq 6$, then $\mathbb{P}(Y = y, X = x) = \mathbb{P}((v_1 = y, v_2 = x) \cup (v_1 = x, v_2 = y))$. There is only 1 outcome in which $v_1 = y, v_2 = x$, and 1 more outcome in which $v_1 = x, v_2 = y$. Since there are 36 equally likely outcomes, we know

$$\mathbb{P}(Y = y, X = x) = \frac{2}{36}$$

for all $1 \leq y < x \leq 6$.
If $x \notin \{1, ..., 6\}$ or $y \notin \{1, ..., 6\}$, then $\mathbb{P}(X = x, Y = y) = 0$. Thus, we know

$$\mathbb{P}(Y = y, X = x) = \begin{cases} \frac{1}{36} \text{ if } 1 \leq y = x \leq 6 \\ \frac{2}{36} \text{ if } 1 \leq y < x \leq 6 \\ 0 \text{ otherwise.} \end{cases}$$

From our calculations for the individual conditional mass functions, we found that, for any $X = i$, there are exactly $2i - 1$ equally likely outcomes in which i is the largest value rolled. This makes sense, as there are 2 outcomes in which one roll is i and the other roll is j for each $j \in \{1, ..., i-1\}$ and 1 outcome in which i is rolled twice, for a total of $2(i-1)+1 = 2i-1$ equally likely outcomes. Since the experiment has 36 equally likely outcomes, we have

$$p_X(x) = \mathbb{P}(X = x) = \begin{cases} \frac{2x-1}{36} \text{ if } 1 \leq x \leq 6 \\ 0 \text{ otherwise.} \end{cases}$$

Plugging in our computed values for $\mathbb{P}(X = x)$ and $\mathbb{P}(Y = y, X = x)$ into (1), we find that the general conditional mass function of Y given X is

$$p_{Y|X}(y|x) = \begin{cases} \frac{1}{2x-1} \text{ if } 1 \leq y = x \leq 6 \\ \frac{2}{2x-1} \text{ if } 1 \leq y < x \leq 6 \\ 0 \text{ otherwise.} \end{cases}$$

For X and Y to be independent, we need

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

for all x, y. We can easily see this statement does not hold under our definition of X and Y. Let $X = 4$, $Y = 5$. Plugging in $x = 4$ into $p_X(x)$, we find

$$p_X(4) = \mathbb{P}(X = 4) = \frac{2(4) - 1}{36} = \frac{7}{36}$$

If $Y = 5$, then at least one roll must be a 5, and the other roll must be no lower than a 5. There are only 3 such outcomes, each of which are equally likely:

$$(v_1, v_2) \in \{(5, 5), (5, 6), (6, 5)\}$$

Since there are 36 total equally likely outcomes, we have

$$\mathbb{P}(Y = 5) = \frac{3}{36}$$

Multiplying these values together, we find

$$\mathbb{P}(Y = 5)\mathbb{P}(X = 4) = \frac{3}{36}\frac{7}{36} = \frac{21}{36^2}$$

However if $X = 4$, the smallest number rolled *cannot* be $Y = 5$, as $5 > 4$. Thus, we know

$$\mathbb{P}(X = 4, Y = 5) = 0 \neq \mathbb{P}(X = 4)\mathbb{P}(Y = 5)$$

This counterexample proves that X and Y are **NOT** independent.
This result makes intuitive sense because, once you know $X = i$, you know that $1 \leq Y \leq i$. Thus, having knowledge about the value of X provides direct knowledge about the value of Y, rendering making the variables dependent on one another. This is consistent with our calculations for $p_{Y|X}(y|x)$, which show that the conditional probability of Y given X takes the value of X as input and thus depends on X.

5. (Ross P7.42) The joint density function of X and Y is given by

$$f(x,y) = \frac{1}{y}e^{-(y+x/y)}, \qquad x > 0, y > 0.$$

Find $E[X], E[Y]$, and show that $E[(X - E[X])(Y - E[Y])] = 1$.

Solution.
First, we will calculate $E[X]$. By the definitions of expected value and the X marginal, we know

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)dydx = \int_0^{\infty}\int_0^{\infty} x\frac{1}{y}e^{-(y+x/y)}dydx$$

We can flip the order of integration to find

$$E[X] = \int_0^{\infty}\int_0^{\infty} \frac{x}{y}e^{-(y+x/y)}dxdy$$

Now, we can integrate by parts with $u = \frac{x}{y}, du = \frac{dx}{y}, dv = e^{-(y+x/y)}dx, v = -ye^{-(y+x/y)}$ to find

$$\begin{aligned}\int_0^{\infty} \frac{x}{y}e^{-(y+x/y)}dx = & \quad -xe^{-(y+x/y)}\Big|_0^{\infty} + \int_0^{\infty} e^{(-y+x/y)}dx \\ = & \quad 0 - 0 + (-ye^{-(y+x/y)})\Big|_0^{\infty} = -0 - (-ye^{-y}) = ye^{-y}\end{aligned}$$

Therefore

$$E[X] = \int_0^{\infty} ye^{-y}dy$$

Integrating by parts with $u = y, du = dy, dv = e^{-y}, v = -e^{-y}$, we find

$$\begin{aligned}E[X] = & \quad -ye^{-y}\Big|_0^{\infty} + \int_0^{\infty} e^{-y}dy \\ = & \quad 0 - 0 + (-e^{-y})\Big|_0^{\infty} = 0 - (-e^0) = e^0 = 1\end{aligned}$$

Thus, the expected value of X is

$$E[X] = 1$$

.

Next, we will calculate $E[Y]$. By the definition of expected value,

$$E[Y] = \int_0^{\infty} y f_Y(y)dy$$

so let's compute $p_Y(y)$ first. By definition

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_0^{\infty} \frac{1}{y}e^{-(y+x/y)}dx$$

If we let $u = y + x/y$, then we have $du = \frac{dx}{y}$, $x = 0 \iff u = y$, and

$$f_Y(y) = \int_y^{\infty} e^{-u}du = -e^{-u}\Big|_y^{\infty} = 0 - (-e^y) = e^{-y}$$

so we have

$$f_Y(y) = \begin{cases} e^{-y} \text{ if } y > 0 \\ 0 \text{ otherwise.} \end{cases}$$

Note that the density function of an exponential random variable with parameter λ is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \text{ if } x > 0 \\ 0 \text{ otherwise.} \end{cases}$$

Since $e^{-y} = 1 \cdot e^{-1 \cdot y}$, we know $Y \sim Exponential(1)$. Thus, Y has expected value

$$E[Y] = \frac{1}{\lambda} = \frac{1}{1} = 1$$

We can also plug in $p_Y(y)$ to the equation for $E[Y]$ to compute directly that

$$E[Y] = \int_0^\infty y e^{-y} dy = E[X] = 1$$

since we already found that $E[X] = \int_0^\infty y e^{-y} dy = 1$.

Plugging $E[Y] = E[X] = 1$ into $E[(X - E[X])(Y - E[Y])]$, we find

$$\begin{aligned} E[(X - E[X])(Y - E[Y])] = & \quad E[(XY - E[X]Y - E[Y]X + E[X]E[Y])] \\ = & \quad E[XY - Y - X + E[X]E[Y]] \\ = & \quad E[XY] - E[Y] - E[X] + 1 = E[XY] - 2 + 1 = E[XY] - 1 \end{aligned}$$

Thus, we just need to compute $E[XY]$. By the definition of expected value, we know

$$\begin{aligned} E[XY] = & \quad \int_{-\infty}^\infty \int_{-\infty}^\infty xy f(x,y) dx dy = \int_0^\infty \int_0^\infty xy \frac{1}{y} e^{-(y+x/y)} dx dy \\ = & \quad \int_0^\infty y \int_0^\infty \frac{x}{y} e^{-(y+x/y)} dx dy \end{aligned}$$

We already computed that

$$\int_0^\infty \frac{x}{y} e^{-(y+x/y)} dx = y e^{-y}$$

when computing $E[X]$, so we know

$$E[XY] = \int_0^\infty y^2 e^{-y} dy$$

Integrating by parts with $u = y^2$, $du = 2y dy$, $dv = e^{-y} dy$, $v = -e^{-y}$, we find

$$\begin{aligned} E[XY] = & \quad -y^2 e^{-y} \Big|_0^\infty + 2 \int_0^\infty y e^{-y} dy \\ = & \quad 0 - 0 + 2(1) = 2 \end{aligned}$$

since we already computed that

$$\int_0^\infty y e^{-y} dy = 1$$

when computing $E[X]$. Thus, we know

$$E[(X - E[X])(Y - E[Y])] = E[XY] - 1 = 2 - 1 = 1$$

which completes the proof that

$$E[(X - E[X])(Y - E[Y])] = 1$$

6. (Ross P7.52) The joint density function of X and Y is given by

$$f(x,y) = \frac{e^{-x/y}e^{-y}}{y}, \qquad 0 < x < \infty, \quad 0 < y < \infty.$$

Compute $E[X^2 \mid Y = y]$.

Solution.
By the definition of expected value and the conditional density function of X given $Y = y$, we know

$$E[X^2 \mid Y = y] = \int_{-\infty}^{\infty} x^2 f_{X|Y}(X|Y=y)dx = \int_{-\infty}^{\infty} x^2 \frac{f(x,y)}{f_Y(y)}dx \tag{1}$$

We are given $f(x,y)$, so we just need to compute $f_Y(y)$. We can compute $f_Y(y)$ directly using its definition.

$$\begin{aligned} f_Y(y) = & \int_{-\infty}^{\infty} f(x,y)dx = \frac{e^{-y}}{y}\int_0^{\infty} e^{-x/y}dx \\ = & \frac{e^{-y}}{y}(-ye^{-x/y})\Big|_{x=0}^{\infty} = \frac{e^{-y}}{y}(0-(-ye^0)) = \frac{e^{-y}}{y}y = e^{-y} \end{aligned}$$

so we have

$$f_Y(y) = \begin{cases} e^{-y} \text{ if } y > 0 \\ 0 \text{ otherwise.} \end{cases}$$

Plugging this into (1), we find

$$E[X^2|Y=y] = \int_0^{\infty} x^2 \frac{\frac{e^{-x/y}e^{-y}}{y}}{e^{-y}}dx = \int_0^{\infty} x^2 \frac{e^{-x/y}}{y}dx = \int_0^{\infty} \frac{x^2}{y}e^{-x/y}dx$$

Integrating by parts with $u = \frac{x^2}{y}$, $du = 2\frac{x}{y}dx$, $dv = e^{-x/y}dx$, $v = -ye^{-x/y}$, we find

$$\begin{aligned} E[X^2|Y=y] = & -x^2e^{-x/y}\Big|_{x=0}^{\infty} + 2\int_0^{\infty} xe^{-x/y}dx \\ = & 0 - 0 + 2\int_0^{\infty} xe^{-x/y}dx = 2\int_0^{\infty} xe^{-x/y}dx \end{aligned}$$

Integrating by parts again with $u = x$, $du = dx$, $dv = e^{-x/y}dx$, $v = -ye^{-x/y}$, we find

$$\begin{aligned} E[X^2|Y=y] = & 2\left(-xye^{-x/y}\Big|_{x=0}^{\infty} + y\int_0^{\infty} e^{-x/y}dx\right) \\ = & 2\left(0 - 0 + y(-ye^{-x/y}|_{x=0}^{\infty})\right) \\ = & 2(0-(-y^2)) = 2y^2 \end{aligned}$$

Thus, the expected value of X^2 given that $Y = y$ is

$$E[X^2|Y=y] = 2y^2$$

7. Suppose X, Y are i.i.d. Exponential(λ) random variables. Compute the density of $X + Y$.

Solution.
We will use the cumulative distribution function of $X + Y$ to solve this problem. By definition, the cumulative distribution function of $X + Y$ is

$$F_{X+Y}(z) = \int\int_{x+y\leq z} f_X(x)f_Y(y)dxdy$$

If we let y go from $-\infty$ to ∞, then we must restrict x from $-\infty$ to $z - y$. Therefore, we have

$$F_{X+Y}(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{z-y} f_X(x)f_Y(y)dxdy = \int_{-\infty}^{\infty} f_Y(y)\left(\int_{-\infty}^{z-y} f_X(x)dx\right)dy$$

By the definition of the cumulative distribution function, we know

$$F_X(z-y) = \int_{-\infty}^{z-y} f_X(x)dx$$

so we can rewrite $F_{X+Y}(z)$ as

$$F_{X+Y}(z) = \int_{-\infty}^{\infty} f_Y(y)F_X(z-y)dy$$

By the definition of the probability density function, we know the density of $X + Y$ is

$$f_{X+Y}(z) = \frac{d}{dz}F_{X+Y}(z)$$

which implies that

$$f_{X+Y}(z) = \frac{d}{dz}\int_{-\infty}^{\infty} f_Y(y)F_X(z-y)dy = \int_{-\infty}^{\infty} f_Y(y)\frac{d}{dz}F_X(z-y)dy = \int_{-\infty}^{\infty} f_Y(y)f_X(z-y)dy$$

Since X and Y are *i.i.d.* Exponential(λ), we know that

$$f_X(x) = f_Y(x) = \begin{cases}\lambda e^{-\lambda x} \text{ if } x > 0\\ 0 \text{ otherwise.}\end{cases}$$

which implies that

$$f_{X+Y}(z) = \int_0^z \lambda e^{-\lambda y}\lambda e^{-\lambda(z-y)}dy$$

The integral's lower bound of 0 stems from the fact that $f_Y(y) = 0$ for all $y \leq 0$. The integral's upper bound of z stems from the fact that $f_X(z-y) = 0$ for all $y \geq z$. We can now directly compute that, for $z > 0$

$$f_{X+Y}(z) = \lambda^2\int_0^z e^{-\lambda(y+z-y)}dy = \lambda^2 e^{-\lambda z}\int_0^z dy = \lambda^2 e^{-\lambda z}y\Big|_{y=0}^{z} = \lambda^2 z e^{-\lambda z}$$

Thus, the density of $X + Y$ is

$$f_{X+Y}(z) = \begin{cases}\lambda^2 z e^{-\lambda z} \text{ if } z > 0\\ 0 \text{ otherwise.}\end{cases}$$

8. (Ross, TE5.13) The median of a continuous random variable having distribution function F is that value m such that $F(m) = \frac{1}{2}$. That is, a random variable is just as likely to be larger than its median as it is to be smaller. Find the median of X if X is

   (a) uniformly distributed over (a, b);

   (b) normal with parameters μ, σ^2;

   (c) exponential with rate λ.

   Solution.

   (a) Let $X \sim ContinuousUniform(a, b)$. Then, by the definition of a Continuous Uniform random variable, X has density function

$$f_X(x) = \begin{cases} \frac{1}{b-a} \text{ if } x \in (a, b) \\ 0 \text{ otherwise.} \end{cases}$$

   Since $F(m) = \frac{1}{2}$, we know

$$F(m) = \int_{-\infty}^{m} f_X(x) = \frac{1}{2}$$

   We can compute directly that

$$F(m) = \int_a^m \frac{1}{b-a} dx = \frac{1}{b-a} x \Big|_a^m = \frac{1}{b-a}(m-a) = \frac{m-a}{b-a}$$

   Thus, we know

$$\frac{1}{2} = \frac{m-a}{b-a} \implies \frac{b-a}{2} = m - a \implies m = \frac{b-a}{2} + a = \frac{b-a+2a}{2} = \frac{b+a}{2}$$

   Thus, the median of $X \sim ContinuousUniform(a, b)$ is

$$m = \frac{b+a}{2}$$

   This makes intuitive sense, as the distribution of X is symmetric about the center of (a, b), which is $\frac{a+b}{2}$.

   (b) Let $N \sim Normal(\mu, \sigma^2)$. Then, by the definition of a Normal random variable, N has density function

$$f_N(x) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

   for all $x \in \mathbb{R}$.
   For $F(m) = \frac{1}{2}$, we need

$$\mathbb{P}(N > m) = \mathbb{P}(N \leq m) = \frac{1}{2}$$

   Since

$$\mathbb{P}(N > m) = 1 - \mathbb{P}(N \leq m) = 1 - F_N(m) = \int_{-\infty}^{\infty} f_N(x)dx - \int_{-\infty}^{m} f_N(x)dx = \int_m^{\infty} f_N(x)dx$$

   we must have

$$\mathbb{P}(N \leq m) = F_N(m) = \int_{-\infty}^{m} f_N(x)dx = \int_m^{\infty} f_N(x)dx = \mathbb{P}(N > m)$$

Using the definition of the density function of a Normal random variable, we find

$$\int_{-\infty}^{m} \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}dx = \int_{m}^{\infty} \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}dx$$

Noting that $\left(\frac{x-\mu}{\sigma}\right)^2$ is part of the power to which e is raised in both integrals, we can let $u = \frac{x-\mu}{\sigma}$, $du = \frac{dx}{\sigma}$ to find

$$\int_{-\infty}^{\frac{m-\mu}{\sigma}} \frac{e^{\frac{-u^2}{2}}}{\sqrt{2\pi}}du = \int_{\frac{m-\mu}{\sigma}}^{\infty} \frac{e^{\frac{-u^2}{2}}}{\sqrt{2\pi}}du$$

Since $\frac{1}{\sqrt{2\pi}}$ is a constant, this implies

$$\int_{-\infty}^{\frac{m-\mu}{\sigma}} e^{\frac{-u^2}{2}}du = \int_{\frac{m-\mu}{\sigma}}^{\infty} e^{\frac{-u^2}{2}}du \qquad (1)$$

Note that, for any $u \in \mathbb{R}$, we have

$$e^{\frac{-(-u)^2}{2}} = e^{\frac{-u^2}{2}} = e^{\frac{-(u)^2}{2}}$$

so $e^{\frac{-u^2}{2}}$ is an even function. For any even function

$$\int_{-\infty}^{0} f(x)dx = \int_{-\infty}^{0} f(-x)dx = \int_{0}^{\infty} f(x)dx$$

so we know (1) will be true if $\frac{m-\mu}{\sigma} = 0$. Thus, we know

$$\frac{m-\mu}{\sigma} = 0 \implies m - \mu = 0 \implies m = \mu$$

We can easily verify (1) holds under $m = \mu$:

$$\int_{-\infty}^{0} e^{\frac{-u^2}{2}}du = \int_{-\infty}^{0} e^{\frac{-(-u)^2}{2}}du = \int_{0}^{\infty} e^{\frac{-u^2}{2}}du$$

Thus, the median of $N \sim Normal(\mu, \sigma^2)$ is

$$m = \mu$$

This also makes intuitive sense, as the distribution of N is symmetric about μ.

(c) Let $E \sim Exponential(\lambda)$. Then, by the definition of an Exponential random variable, E has density function

$$f_E(x) = \begin{cases} \lambda e^{-\lambda x} \text{ if } x > 0 \\ 0 \text{ otherwise.} \end{cases}$$

Since $F(m) = \frac{1}{2}$, we know

$$F(m) = \int_{-\infty}^{m} f_E(x)dx = \int_{0}^{m} \lambda e^{-\lambda x}dx = \frac{1}{2}$$

We can compute directly that

$$\frac{1}{2} = F(m) = \int_{0}^{m} \lambda e^{-\lambda x}dx = -e^{-\lambda x}\Big|_{0}^{m} = -e^{-\lambda m} - (-e^0) = 1 - e^{-\lambda m}$$

This implies

$$\frac{-1}{2} = -e^{-\lambda m} \implies \frac{1}{2} = e^{-\lambda m} \implies ln(\frac{1}{2}) = -\lambda m \implies \frac{-ln(\frac{1}{2})}{\lambda} = \frac{ln(2)}{\lambda} = m$$

Thus, the median of $E \sim Exponential(\lambda)$ is

$$m = \frac{ln(2)}{\lambda}$$

This also makes intuitive sense, as $\frac{ln(2)}{\lambda} < \frac{1}{\lambda} = E[E]$, and E's distribution is skewed to the right, so the median should be less than the mean.

Assignment 13

Math 407 (Swanson) – Spring 2023
Homework 1
Due Friday 1/13, 11:59pm

Name: Emerson Kahle Section: 39981

- You must upload your solutions to Gradescope as **one single, high-quality PDF**. You can convert paper-based work to a high-quality PDF using a scanning app for mobile devices, such as Adobe Scan (free, available for iOS and Android, can do multiple pages) or many others. If necessary, you can combine or merge multiple PDF's into a single PDF using a variety of services, such as Adobe Acrobat's cloud-based merge tool.
- After you upload, you must match each question with its corresponding page using Gradescope's interface. This allows graders to spend more time giving you feedback instead of hunting through submissions.
- Answers without supporting work will receive no credit. Show your work.
- You are encouraged to work together on homework, but **you must write up your solutions separately in your own words**. Copying from your fellow students or other sources is a serious academic integrity violation. In particular, you may not use "tutoring" services which simply provide answers.
- You are encouraged to typeset your solutions in LaTeX. Source code has been provided on Blackboard. Overleaf is a popular cloud-based editor.
- Problem numbers refer to the course textbook, though the problems may have been modified significantly.

1. (Ross, P5.16) The annual rainfall (in inches) in a certain region is normally distributed with $\mu = 40$ and $\sigma = 4$. What is the probability that starting with this year, it will take more than 10 years before a year occurs having a rainfall of more than 50 inches? What assumptions are you making?

Solution.
Let X_i = the annual rainfall in inches for the ith year, starting from this year.
In order to arrive at a conclusive answer, we assume that the annual rainfall in the ith year is independent from the rainfall in the jth year for all $i \neq j$.
Let Y = the number of years, starting from this year, before a year occurs having a rainfall of more than 50 inches. Then we want to find

$$\mathbb{P}(Y > 10)$$

Let $Z_i = \begin{cases} 1 \text{ if } X_i > 50 \\ 0 \text{ otherwise.} \end{cases}$
Then the Z_is are *i.i.d. Bernoulli*(p) random variables, where

$$p = \mathbb{P}(X_i > 50)$$

Since the X_is are *i.i.d. Normal*$(40, 16)$, we can compute that

$$\mathbb{P}(X_i > 50) = \mathbb{P}(\frac{X_i - 40}{4} > \frac{50 - 40}{4}) = \mathbb{P}(\frac{X_i - 40}{4} > 2.5) = 1 - \mathbb{P}(\frac{X_i - 40}{4} \leq 2.5) = 1 - \phi(2.5)$$

since $\frac{X_i - 40}{4}$ is a standardized normal random variable. We can approximate that

$$\mathbb{P}(X_i > 50) \approx 1 - 0.99379 = 0.00621$$

Thus, we have that the Z_is are *i.i.d. Bernoulli*(p) where $p = 0.00621$. Since Y is simply the number of Z_is until the first 'success' ($Z_i = 1$), we have that $Y \sim$ *Geometric*(p). Thus, Y has probability mass function

$$p_Y(y) = \begin{cases} (1-p)^{y-1}p \text{ if } y \in \mathbb{N} \\ 0 \text{ otherwise.} \end{cases}$$

so we can directly compute that

$$\begin{aligned} \mathbb{P}(Y > 10) = & \quad 1 - P(Y \leq 10) = 1 - \sum_{i=1}^{10}(1-p)^{i-1}p = 1 - p\sum_{i=0}^{9}(1-p)^i \\ = & \quad 1 - p\frac{1-(1-p)^{10}}{1-(1-p)} = 1 - (1-(1-p)^{10}) = (1-p)^{10} \\ = & \quad (0.99379)^{10} \approx 0.9396 = 93.96\% \end{aligned}$$

Thus, there is approximately a 93.96% chance it will take more than 10 years before a year occurs having a rainfall of more than 50 years, assuming the annual rainfalls are mutually independent.

2. (Ross, P5.23) One thousand independent rolls of a fair die will be made. Compute an approximation to the probability that the number 6 will appear between 150 and 200 times (i.e. in $[150, 200]$). If the number 6 appears exactly 200 times, find the probability that the number 5 will appear less than 150 times.

   (Note: here it is best to use the *continuity correction*, which replaces the discrete probability $P(X = i)$ with $P(i - 1/2 < X < i + 1/2)$. This is, however, typically only a small improvement.)

   Solution.
   Let $X_i = \begin{cases} 1 \text{ if the } i\text{th roll is a 6} \\ 0 \text{ otherwise.} \end{cases}$
   Then the 1000 independent dice rolls correspond to 1000 *i.i.d.* $X_i \sim Bernoulli(\frac{1}{6})$ random variables. Let $Y = \sum_{i=1}^{1000} X_i$ = the number of 6's rolled over the 1000 independent dice rolls. Then $Y \sim Binomial(1000, \frac{1}{6})$, so we know

$$\mu_Y = \frac{1000}{6} \qquad \sigma_Y^2 = \frac{1}{6}\frac{5}{6}1000 = \frac{5000}{36} = \frac{1250}{9}$$

   Using the Normal approximation to the Binomial distribution since $n = 1000$ is large, and applying the *continuity correction*, we find

$$\mathbb{P}(150 \le Y \le 200) \approx \mathbb{P}(149.5 \le Z \le 200.5)$$

   where $Z \sim Normal(\mu_Y, \sigma_Y^2)$. Thus, we can compute that

$$\begin{aligned} \mathbb{P}(150 \le Y \le 200) \approx & \quad \mathbb{P}\left(\frac{149.5 - \frac{1000}{6}}{\sqrt{\frac{1250}{9}}} \le \frac{Z - \frac{1000}{6}}{\sqrt{\frac{1250}{9}}} \le \frac{200.5 - \frac{1000}{6}}{\sqrt{\frac{1250}{9}}}\right) \\ = & \quad \mathbb{P}\left(-1.457 \le \frac{Z - \frac{1000}{6}}{\sqrt{\frac{1250}{9}}} \le 2.871\right) = \phi(2.871) - \phi(-1.457) \end{aligned}$$

   since $\frac{Z - \frac{1000}{6}}{\sqrt{\frac{1250}{9}}}$ is a standardized normal random variable. Thus, we can approximate that

$$\mathbb{P}(150 \le Y \le 200) \approx 0.9980 - 0.0726 = 0.9254 = 92.54\%$$

   Thus, the probability that the number 6 will appear between 150 and 200 times in 1000 independent rolls of a fair die is approximately 92.54%.

   If we are given that the number 6 appears exactly 200 times, then we know the other 800 independent dice rolls only have values from 1 to 5. Let $A_i = \begin{cases} 1 \text{ if the } i\text{th roll is a 5} \\ 0 \text{ otherwise.} \end{cases}$ Then the A_i's are 800 *i.i.d.* $Bernoulli(\frac{1}{5})$ random variables.
   Let B = the number of 5s, given that there are exactly 200 6s. Then $B = \sum_{i=1}^{800} A_i$, so $B \sim Binomial(800, \frac{1}{5})$. Thus, B has mean $\mu_B = \frac{800}{5} = 160$ and variance $\sigma_B^2 = \frac{800*4*1}{5*5} = \frac{3200}{25} = 128$. Using the Normal approximation for the Binomial distribution since $n = 800$ is large, and applying the *continuity correction*, we find the probability that the number 5 appears less than 150 times, given that the number 6 appears exactly 200 times, is

$$\begin{aligned} \mathbb{P}(B < 150) = & \quad \mathbb{P}(B \le 149) \approx \mathbb{P}(C \le 149.5) = \mathbb{P}\left(\frac{C - 160}{\sqrt{128}} \le \frac{149.5 - 160}{\sqrt{128}}\right) \\ = & \quad \mathbb{P}\left(\frac{C - 160}{\sqrt{128}} \le -0.9281\right) = \phi(-0.9281) \end{aligned}$$

where $C \sim Normal(160, 128)$ since $\frac{C-160}{\sqrt{128}}$ is a standardized normal random variable. Thus, we can approximate that

$$\mathbb{P}(B < 150) \approx 0.1767 = 17.67\%$$

Thus, the probability that the number 5 appears less than 150 times given that the number 6 appears exactly 200 times is approximately 17.67%.

3. (Ross, P8.5) Fifty numbers are rounded off to the nearest integer and then summed. If the individual round-off errors are uniformly distributed over $(-0.5, 0.5)$, approximate the probability that the resultant sum differs from the exact sum by more than 3.

Solution.
Let X_i = the round-off error of the ith number. Then $\{X_1, \cdots, X_n\}$ are *i.i.d.* $ContinuousUniform(-0.5, 0.5)$. Thus, they all have mean $\mu_X = \frac{-0.5+0.5}{2} = \frac{0}{2} = 0$ and variance $\sigma_X^2 = \frac{(0.5-(-0.5))^2}{12} = \frac{1}{12}$.
Let Y = the difference between the exact and rounded sums of the fifty numbers. Then

$$Y = \sum_{i=1}^{50} X_i$$

so Y has mean $\mu_Y = 50 \cdot \mu_X = 0$ and variance $\sigma_Y^2 = 50\sigma_X^2 = \frac{50}{12} = \frac{25}{6}$. Thus, we can apply the Central Limit Theorem to find

$$\begin{aligned}\mathbb{P}((Y < -3) \cup (Y > 3)) = &\quad 1 - \mathbb{P}(-3 \leq Y \leq 3) = 1 - \mathbb{P}(\frac{-3}{\sqrt{\frac{25}{6}}} \leq \frac{Y}{\sqrt{\frac{25}{6}}} \leq \frac{3}{\sqrt{\frac{25}{6}}}) \\ \approx &\quad 1 - \mathbb{P}(\frac{-3\sqrt{6}}{5} \leq \frac{Z\sqrt{6}}{5} \leq \frac{3\sqrt{6}}{5}) = 1 - (\phi(\frac{3\sqrt{6}}{5}) - \phi(\frac{-3\sqrt{6}}{5}))\end{aligned}$$

where $Z \sim Normal(0, \frac{25}{6})$ since $\frac{Z\sqrt{6}}{5}$ is a standardized normal random variable. Thus we can approximate that

$$\mathbb{P}((Y < -3) \cup (Y > 3)) \approx 1 - (0.9292 - 0.0708) = 1 - 0.8584 = 0.1416 = 14.16\%$$

Thus, the probability that the rounded and exact sums of the fifty numbers differ by more than 3 is approximately 14.16%.

Note: This answer assumes that the question is asking about the *absolute* difference between the rounded and exact sums. If the question is asking specifically for the probability that the rounded sum *exceeds* the exact sum by more than 3, we have

$$\mathbb{P}(Y > 3) = 1 - \mathbb{P}(Y \leq 3) \approx 1 - \mathbb{P}(\frac{Z\sqrt{6}}{5} > \frac{3\sqrt{6}}{5}) = 1 - \phi(\frac{3\sqrt{6}}{5}) \approx 1 - 0.9292 = 0.0708 = 7.08\%$$

In this case, the probability that the rounded sum *exceeds* the exact sum by more than 3 is approximately 7.08%.

4. (Ross, P8.15) An insurance company has 10,000 automobile policyholders. The expected yearly claim per policyholder is \$240, with a standard deviation of \$800. Approximate the probability that the total yearly claim exceeds \$2.7 million.

Solution.
Let X_i = the yearly claim of the ith policyholder.
Let Y = the combined yearly claim of all 10,000 automobile policyholders. Then, assuming the X_is are *i.i.d.* with $\mu_X = 240$ and $\sigma_X = 800$, we have

$$\mu_Y = 10,000 \cdot 240 = 2,400,000 \qquad \sigma_Y^2 = 10,000 \cdot (800)^2 = 10,000 \cdot 640,000 = 6,400,000,000 \qquad \sigma_Y = \sqrt{6,400,000,000} = 80,000$$

Thus we can apply the Central Limit Theorem with $Z \sim Normal(2,400,000, 6,400,000,000)$ to find

$$\begin{aligned}
\mathbb{P}(Y > 2,700,000) \approx & \quad \mathbb{P}(Z > 2,700,000) = \mathbb{P}\left(\frac{Z - 2,400,000}{80,000} > \frac{2,700,000 - 2,400,000}{80,000}\right) \\
= & \quad \mathbb{P}\left(\frac{Z - 2,400,000}{80,000} > 3.75\right) = 1 - \mathbb{P}\left(\frac{Z - 2,400,000}{80,000} \leq 3.75\right) \\
= & \quad 1 - \phi(3.75) \approx 1 - 0.99991 = 0.00009 = 0.009\%
\end{aligned}$$

since $\frac{Z-2,400,000}{80,000}$ is a standardized normal random variable.
Thus, the probability that the total yearly claim exceeds \$2.7 million is approximately 0.009%.

5. (Ross, TE5.31) Find the probability density function of $Y = e^X$ when X is normally distributed with parameters μ and σ^2. The random variable Y is said to have a *lognormal distribution* (since $\log Y$ has a normal distribution) with parameters μ and σ^2.

Solution.
We will use the fact that

$$\frac{d}{dx}F_Y(x) = \frac{d}{dx}\mathbb{P}(Y \leq x) = f_Y(x)$$

We can compute directly that

$$F_Y(x) = \mathbb{P}(Y \leq x) = \mathbb{P}(e^X \leq x) = \mathbb{P}(X \leq ln(x)) = F_X(ln(x))$$

This implies Y has probability density function

$$f_Y(x) = \frac{d}{dx}F_X(ln(x)) = \frac{1}{x}f_X(ln(x))$$

Since $X \sim Normal(\mu, \sigma^2)$, we know X has density function

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for all $x \in \mathbb{R}$. This implies

$$f_Y(x) = \frac{1}{x}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}} = \frac{e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}}}{x\sigma\sqrt{2\pi}}$$

for all $x \in \mathbb{R}$. Thus, if $X \sim Normal(\mu, \sigma^2)$, then $Y = e^X$ has probability density function

$$f_Y(x) = \frac{e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}}}{x\sigma\sqrt{2\pi}}$$

for all $x \in \mathbb{R}$.

6. (Ross, P7.38) Suppose X and Y have the following joint probability mass function.

$$\begin{aligned} p(1,1) &= 0.10, \quad p(1,2) = 0.12, \quad p(1,3) = 0.16 \\ p(2,1) &= 0.08, \quad p(2,2) = 0.12, \quad p(2,3) = 0.10 \\ p(3,1) &= 0.06, \quad p(3,2) = 0.06, \quad p(3,3) = 0.20 \end{aligned}$$

(a) Find $E[X]$ and $E[Y]$.

(b) Find $\text{Var}(X)$ and $\text{Var}(Y)$.

(c) Find $\text{Cov}(X, Y)$.

(d) Find the correlation between X and Y.

Solution.

(a) First, we will find $E[X]$. By the definition of the expected value of a discrete random variable, we know

$$E[X] = \sum_{x=1}^{3} x\mathbb{P}(X = x) \qquad (1)$$

We can easily compute that

$$\begin{aligned} \mathbb{P}(X = 1) = \quad & \mathbb{P}((X = 1, Y = 1) \cup (X = 1, Y = 2) \cup (X = 1, Y = 3)) \\ = \quad & \mathbb{P}(X = 1, Y = 1) + \mathbb{P}(X = 1, Y = 2) + \mathbb{P}(X = 1, Y = 3) \\ = \quad & p(1,1) + p(1,2) + p(1,3) = 0.10 + 0.12 + 0.16 = 0.38 \end{aligned}$$

Similarly, we have

$$\begin{aligned} \mathbb{P}(X = 2) = \quad & \mathbb{P}((X = 2, Y = 1) \cup (X = 2, Y = 2) \cup (X = 2, Y = 3)) \\ = \quad & \mathbb{P}(X = 2, Y = 1) + \mathbb{P}(X = 2, Y = 2) + \mathbb{P}(X = 2, Y = 3) \\ = \quad & p(2,1) + p(2,2) + p(2,3) = 0.08 + 0.12 + 0.10 = 0.30 \end{aligned}$$

and

$$\begin{aligned} \mathbb{P}(X = 3) = \quad & \mathbb{P}((X = 3, Y = 1) \cup (X = 3, Y = 2) \cup (X = 3, Y = 3)) \\ = \quad & \mathbb{P}(X = 3, Y = 1) + \mathbb{P}(X = 3, Y = 2) + \mathbb{P}(X = 3, Y = 3) \\ = \quad & p(3,1) + p(3,2) + p(3,3) = 0.06 + 0.06 + 0.20 = 0.32 \end{aligned}$$

Plugging these values into (1), we find

$$E[X] = 1 \cdot 0.38 + 2 \cdot 0.30 + 3 \cdot 0.32 = 0.38 + 0.60 + 0.96 = 1.94$$

Now, we can find $E[Y]$. By the definition of expected value, we know

$$E[Y] = \sum_{y=1}^{3} y\mathbb{P}(Y = y) \qquad (2)$$

We can easily compute that

$$\begin{aligned} \mathbb{P}(Y = 1) = \quad & \mathbb{P}((X = 1, Y = 1) \cup (X = 2, Y = 1) \cup (X = 3, Y = 1)) \\ = \quad & \mathbb{P}(X = 1, Y = 1) + \mathbb{P}(X = 2, Y = 1) + \mathbb{P}(X = 3, Y = 1) \\ = \quad & p(1,1) + p(2,1) + p(3,1) = 0.10 + 0.08 + 0.06 = 0.24 \end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathbb{P}(Y=2) = &\quad \mathbb{P}((X=1,Y=2)\cup(X=2,Y=2)\cup(X=3,Y=2))\\ = &\quad \mathbb{P}(X=1,Y=2)+\mathbb{P}(X=2,Y=2)+\mathbb{P}(X=3,Y=2)\\ = &\quad p(1,2)+p(2,2)+p(3,2)=0.12+0.12+0.06=0.30\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(Y=3) = &\quad \mathbb{P}((X=1,Y=3)\cup(X=2,Y=3)\cup(X=3,Y=3))\\ = &\quad \mathbb{P}(X=1,Y=3)+\mathbb{P}(X=2,Y=3)+\mathbb{P}(X=3,Y=3)\\ = &\quad p(1,3)+p(2,3)+p(3,3)=0.16+0.10+0.20=0.46\end{aligned}$$

Plugging these values into (2), we find

$$E[Y] = 1\cdot 0.24 + 2\cdot 0.30 + 3\cdot 0.46 = 2.22$$

(b) We know that

$$Var(X) = E[X^2] - E[X]^2 \qquad (3)$$

We already computed $E[X]$, so we just need to compute $E[X^2]$. By the definition of the second raw moment, we know

$$E[X^2] = \sum_{x=1}^{3} x^2\mathbb{P}(X=x) = 1\cdot 0.38 + 4\cdot 0.30 + 9\cdot 0.32 = 4.46$$

Thus, we can directly compute that

$$Var(X) = 4.46 - 1.94^2 = 4.46 - 3.7636 = 0.6964$$

Similarly, we know that

$$Var(Y) = E[Y^2] - E[Y]^2 \qquad (4)$$

and since we already computed $E[Y]$, we just need to find $E[Y^2]$. By the definition of the second raw moment, we know

$$E[Y^2] = \sum_{y=1}^{3} y^2\mathbb{P}(Y=y) = 1\cdot 0.24 + 4\cdot 0.30 + 9\cdot 0.46 = 5.58$$

Plugging this into (4) yields

$$Var(Y) = 5.58 - 2.22^2 = 5.58 - 4.9284 = 0.6516$$

(c) By the definition of Covariance, we know

$$Cov(X,Y) = E[XY] - E[X]E[Y] \qquad (5)$$

We already computed $E[X]$ and $E[Y]$, so we just need to compute $E[XY]$. To compute this, we can simply sum over all possible combinations of X and Y, multiplying xy by $p(x,y)$ for each combination. We find

$$\begin{aligned}E[XY] = &\quad \sum_{x=1}^{3}\sum_{y=1}^{3} xyp(x,y)\\ = &\quad (1\cdot 1\cdot 0.10)+(1\cdot 2\cdot 0.12)+(1\cdot 3\cdot 0.16)+(2\cdot 1\cdot 0.08)+(2\cdot 2\cdot 0.12)+(2\cdot 3\cdot 0.10)\\ &\quad +(3\cdot 1\cdot 0.06)+(3\cdot 2\cdot 0.06)+(3\cdot 3\cdot 0.20)\\ = &\quad 0.10+0.24+0.48+0.16+0.48+0.60+0.18+0.36+1.8\\ = &\quad 4.4\end{aligned}$$

Plugging this into (5), we find

$$Cov(X,Y) = 4.4 - (1.94)(2.22) = 4.4 - 4.3068 = 0.0932$$

Thus, the Covariance of X and Y is 0.0932.

(d) By definition, the correlation between X and Y is

$$Corr(X,Y) = \frac{Cov(X,Y)}{SD(X)SD(Y)} \qquad (6)$$

We already calculated $Cov(X,Y)$ in part (c), and we know that $SD(X) = \sqrt{Var(X)} = \sqrt{0.6964}$ and $SD(Y) = \sqrt{Var(Y)} = \sqrt{0.6516}$. Plugging these values into (6), we can easily compute that

$$Corr(X,Y) = \frac{0.0932}{\sqrt{0.6964}\sqrt{0.6516}} \approx 0.1384$$

Thus, the correlation between X and Y is approximately 0.1384, so X and Y have a weak positive correlation.

7. (Ross, P7.39) Suppose that 2 balls are randomly removed from an urn containing n red and m blue balls. Let $X_i = 1$ if the ith ball removed is red, and let it be 0 otherwise, $i = 1, 2$.

   (a) Do you think that $\text{Cov}(X_1, X_2)$ is negative, zero, or positive?

   (b) Validate your answer to part (a).

   Solution.

   (a) I think that $Cov(X_1, X_2)$ is *negative*. If the first ball picked is *not* red (i.e. $X_1 = 0$), then a higher proportion of the remaining balls will be red, so there should be a higher probability that the second ball picked is red (i.e. $X_2 = 1$). Similarly, if the first ball picked *is* red (i.e. $X_1 = 1$), then a lower proportion of the remaining balls will be red, so there should be a lower probability that the second ball picked is red (i.e. $X_2 = 1$).

   (b) By definition,
   $$Cov(X_1, X_2) = E[X_1 X_2] - E[X_1]E[X_2]$$
   We can easily see that $X_1 \sim Bernoulli(\frac{n}{n+m})$, so we know
   $$E[X_1] = \mathbb{P}(X_1 = 1) = \frac{n}{n+m}$$
   We can use the Law of Total Probability to compute that
   $$\begin{aligned} E[X_2] = \quad & \mathbb{P}(X_2 = 1) = \mathbb{P}(X_2 = 1|X_1 = 1)\mathbb{P}(X_1 = 1) + \mathbb{P}(X_2 = 1|X_1 = 0)\mathbb{P}(X_1 = 0) \\ = \quad & \frac{n-1}{n+m-1}\frac{n}{n+m} + \frac{n}{n+m-1}\frac{m}{n+m} \\ = \quad & \frac{n(m+n-1)}{(n+m)(n+m-1)} = \frac{n}{n+m} \end{aligned}$$
   so $X_2 \sim Bernoulli(\frac{n}{n+m})$ as well. However, as X_1 and X_2 are not independent, we find
   $$E[X_1 X_2] = \mathbb{P}((X_1 = 1) \cap (X_2 = 1)) = \frac{n}{n+m}\frac{n-1}{n+m-1}$$
   This allows us to compute that
   $$Cov(X_1, X_2) = \frac{n}{n+m}\frac{n-1}{n+m-1} - \frac{n^2}{(n+m)^2}$$
   We want to show
   $$Cov(X_1, X_2) < 0$$
   Note that
   $$\begin{aligned} Cov(X_1, X_2) = \quad & \frac{n}{n+m}\left(\frac{n-1}{n+m-1} - \frac{n}{n+m}\right) \\ = \quad & \frac{n}{n+m}\left(\frac{(n-1)(n+m) - n(n+m-1)}{(n+m)(n+m-1)}\right) \\ = \quad & \frac{n}{n+m}\left(\frac{n^2 - n + nm - m - n^2 - nm - n}{(n+m)(n+m-1)}\right) \\ = \quad & \frac{n}{n+m}\left(\frac{-2n-m}{(n+m)(n+m-1)}\right) \end{aligned}$$
   Assuming both n and m are positive, we have
   $$Cov(X_1, X_2) < 0$$
   since $-2n-m < 0$. This completes the verification of the intuition from part (a) that $Cov(X_1, X_2) < 0$.

8. (Ross, P7.49) Consider a graph having n vertices labeled $1, 2, \ldots, n$, and suppose that, between each of the $\binom{n}{2}$ pairs of distinct vertices, an edge is independently present with probability p. The degree of vertex i, designated D_i, is the number of edges that have vertex i as one of their vertices.

   (a) What is the distribution of D_i?

   (b) Find $\rho(D_i, D_j)$, the correlation between D_i and D_j.

   (Incidentally, this method of creating random graphs is called an *Erdös–Rényi model.*)

   Solution.

   (a) Note that, for each vertex i, there are $n-1$ potential edges incident to i. Pick a vertex i.

   If we let $X_{i,j} = \begin{cases} 1 \text{ if there is an edge between vertices } i \text{ and } j \\ 0 \text{ otherwise.} \end{cases}$ for all $j \neq i$

   Then the $n-1$ X_{ij}'s are *i.i.d. Bernoulli*(p) random variables, and the degree of vertex i is

$$D_i = \sum_{1 \leq j \neq i \leq n} X_{i,j}$$

   Since D_i is the sum of $n-1$ *i.i.d. Bernoulli*(p) random variables, we know $D_i \sim Binomial(n-1, p)$. Thus, the distribution of the degree of vertex i is Binomial with parameters $n-1$ and p.

   (b) By definition, the correlation between D_i and D_j is

$$\rho(D_i, D_j) = Corr(D_i, D_j) = \frac{Cov(D_i, D_j)}{SD(D_i)SD(D_j)} \qquad (1)$$

   Since D_i and D_j are $Binomial(n-1, p)$ random variables, we know

$$E[D_i] = E[D_j] = (n-1)p \qquad Var(D_i) = Var(D_j) = (n-1)p(1-p)$$

   so we just need to compute $Cov(D_i, D_j)$. Note that, if we redefine

$$X_{a,b} = \begin{cases} 1 \text{ if there is an edge between vertices } a \text{ and } b \\ 0 \text{ otherwise.} \end{cases}$$ for all $b \neq a$

   we have

$$Cov(D_i, D_j) = Cov(\sum_{1 \leq x \neq i \leq n} X_{i,x}, \sum_{1 \leq y \neq j \leq n} X_{j,y}) = \sum_{1 \leq x \neq i \leq} \sum_{1 \leq y \neq j \leq n} Cov(X_{i,x}, X_{j,y})$$

   Note: For all $x \neq j$ or $y \neq i$, $X_{i,x}$ and $X_{j,y}$ are different edges, so they have independent probability p of existing, and are thus independent random variables. Therefore, we have

$$Cov(X_{i,x}, X_{j,y}) = E[X_{i,x}X_{j,y}] - E[X_{i,x}]E[X_{j,y}] = E[X_{i,x}]E[X_{j,y}] - E[X_{i,x}]E[X_{j,y}] = 0$$

   for all $x \neq j$ or $y \neq i$. This implies

$$Cov(D_i, D_j) = \sum_{1 \leq x \neq i \leq} \sum_{1 \leq y \neq j \leq n} Cov(X_{i,x}, X_{j,y}) = Cov(X_{i,j}, X_{j,i}) = E[X_{i,j}X_{j,i}] - E[X_{i,j}]E[X_{j,i}]$$

   We already know that

$$E[X_{i,j}] = E[X_{j,i}] = p$$

   Note that

$$E[X_{i,j}X_{j,i}] = \begin{cases} 1 \text{ if there is an edge between vertices } i \text{ and } j \\ 0 \text{ otherwise.} \end{cases}$$

so $X_{i,j}X_{j,i}$ is also a $Bernoulli(p)$ random variable. Thus, we know $E[X_{i,j}X_{j,i}] = p$, which implies

$$Cov(D_i, D_j) = p - p \cdot p = p - p^2 = p(1-p)$$

Plugging this into (1), along with $SD(D_i) = SD(D_j) = \sqrt{Var(D_i)} = \sqrt{(n-1)p(1-p)}$, we find

$$\rho(D_i, D_j) = Corr(D_i, D_j) = \frac{p(1-p)}{\sqrt{(n-1)p(1-p)}\sqrt{(n-1)p(1-p)}} = \frac{p(1-p)}{(n-1)p(1-p)} = \frac{1}{n-1}$$

Thus, the correlation between D_i and D_j is $\frac{1}{n-1}$.

Assignment 14

Math 407 (Swanson) – Spring 2023
Homework 1
Due Friday 1/13, 11:59pm

Name: Emerson Kahle Section: 39981

- You must upload your solutions to Gradescope as **one single, high-quality PDF**. You can convert paper-based work to a high-quality PDF using a scanning app for mobile devices, such as Adobe Scan (free, available for iOS and Android, can do multiple pages) or many others. If necessary, you can combine or merge multiple PDF's into a single PDF using a variety of services, such as Adobe Acrobat's cloud-based merge tool.
- After you upload, you must match each question with its corresponding page using Gradescope's interface. This allows graders to spend more time giving you feedback instead of hunting through submissions.
- Answers without supporting work will receive no credit. Show your work.
- You are encouraged to work together on homework, but **you must write up your solutions separately in your own words**. Copying from your fellow students or other sources is a serious academic integrity violation. In particular, you may not use "tutoring" services which simply provide answers.
- You are encouraged to typeset your solutions in LaTeX. Source code has been provided on Blackboard. Overleaf is a popular cloud-based editor.
- Problem numbers refer to the course textbook, though the problems may have been modified significantly.

1. Let X, Y be jointly continuous random variables such that $X - Y$ and $X + Y$ are i.i.d. standard normal random variables. Show that (X, Y) is a bivariate normal random variable. Explicitly compute the covariance matrix Σ.

Solution.
By definition, we know (X, Y) is a bivariate normal random variable $\iff \exists X_1, X_2$ i.i.d. $Normal(0, 1)$ random variables such that

$$X = a_{11}X_1 + a_{12}X_2 + \mu_X$$
$$Y = a_{21}X_1 + a_{22}X_2 + \mu_Y$$

where $a_{11}, a_{12}, a_{21}, a_{22} \in \mathbb{R}$.
We know $X - Y$ and $X + Y$ are i.i.d. $Normal(0, 1)$ random variables, so let $X_1 = (X + Y)$ and $X_2 = (X - Y)$. We can quickly see that

$$\begin{aligned} X = & \quad \frac{1}{2}X_1 + \frac{1}{2}X_2 + \mu_X = \frac{X+Y}{2} + \frac{X-Y}{2} + 0 \\ = & \quad \frac{X}{2} + \frac{X}{2} + \frac{Y}{2} - \frac{Y}{2} = X \end{aligned} \tag{1}$$

Similarly, we can clearly see that

$$\begin{aligned} Y = & \quad \frac{1}{2}X_1 + \frac{-1}{2}X_2 + \mu_Y = \frac{X+Y}{2} - \frac{X-Y}{2} + 0 \\ = & \quad \frac{X}{2} - \frac{X}{2} + \frac{Y}{2} + \frac{Y}{2} = Y \end{aligned} \tag{2}$$

Combining (1) and (2), we find $a_{11} = \frac{1}{2} = a_{12} = a_{21}$ and $a_{22} = \frac{-1}{2}$. Thus, we have found real numbers $a_{11}, a_{12}, a_{21}, a_{22}$ and standard normal random variables X_1, X_2 such that

$$X = a_{11}X_1 + a_{12}X_2 + \mu_X$$
$$Y = a_{21}X_1 + a_{22}X_2 + \mu_Y$$

which proves that (X, Y) is a bivariate normal random variable by the definition of the bivariate normal random variable.

To compute the covariance matrix $\sum$, we can apply the property that

$$\sum = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} a_{11}^2 + a_{12}^2 & a_{11}a_{21} + a_{12}a_{22} \\ a_{11}a_{21} + a_{12}a_{22} & a_{21}^2 + a_{22}^2 \end{bmatrix}$$

to find that the covariance matrix of (X, Y) is

$$\begin{aligned} \sum = & \quad \begin{bmatrix} (\frac{1}{2})^2 + (\frac{1}{2})^2 & (\frac{1}{2})(\frac{1}{2}) + (\frac{1}{2})(-\frac{1}{2}) \\ (\frac{1}{2})(-\frac{1}{2}) & (\frac{1}{2})^2 + (-\frac{1}{2})^2 \end{bmatrix} \\ = & \quad \begin{bmatrix} \frac{1}{4} + \frac{1}{4} & \frac{1}{4} - \frac{1}{4} \\ \frac{1}{4} - \frac{1}{4} & \frac{1}{4} + \frac{1}{4} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \end{aligned}$$

2. Recall that a bivariate normal random variable (Y_1, Y_2) is determined by the 5 parameters $\mu_i = E[Y_i]$, $\sigma_i^2 = \text{Var}(Y_i)$, $\rho = \text{Cov}(Y_1, Y_2)/(\sigma_1\sigma_2)$.

   (a) Let

$$Y_1 = \sigma_1 X_1 + \mu_1$$
$$Y_2 = \rho\sigma_2 X_1 + \sigma_2\sqrt{1-\rho^2}X_2 + \mu_2$$

   where X_1, X_2 are i.i.d. standard normal random variables. Show that (Y_1, Y_2) have the 5 parameters above.
   (Aside for those who have seen linear algebra: this is related to the *Cholesky decomposition* of the covariance matrix Σ. Here $\Sigma = AA^T$ where $A = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{pmatrix}$.)

   (b) Recall HW 11, Exercise 7, where you used a routine `Uniform(a, b)` to construct a new routine to sample from exponential distributions. In this problem, use (a) and `Uniform(a, b)` to write another routine `BivariateNormal(mu1, mu2, sigma1, sigma2, rho)` that samples from a bivariate normal random variable with given parameters $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho$. You may assume there is another routine `InversePhi(t)` which returns the unique value x such that $\Phi(x) = t$ where Φ is the CDF of the standard normal.

Solution.

   (a) Since we are given

$$A = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{pmatrix}$$

   we can easily compute that the covariance matrix of (Y_1, Y_2) is

$$\begin{aligned}
\sum = & \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = AA^T = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{pmatrix}\begin{pmatrix} \sigma_1 & \rho\sigma_2 \\ 0 & \sigma_2\sqrt{1-\rho^2} \end{pmatrix} \\
= & \begin{pmatrix} \sigma_1(\sigma_1)+0(0) & \sigma_1\rho\sigma_2 + 0\cdot\sigma_2\sqrt{1-\rho^2} \\ \rho\sigma_2\sigma_1 + \sigma_2\sqrt{1-\rho^2}\cdot 0 & \rho\sigma_2\cdot\rho\sigma_2 + (\sigma_2\sqrt{1-\rho^2})^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \rho^2\sigma_2^2 + \sigma_2^2(1-\rho^2) \end{pmatrix} \\
= & \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \rho^2\sigma_2^2 + \sigma_2^2 - \rho^2\sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}
\end{aligned}$$

   This completes the verification that (X, Y) has the parameters $E[Y_1] = \mu_1$, $E[Y_2] = \mu_2$, $Var(Y_1) = \sigma_1^2$, $Var(Y_2) = \sigma_2^2$, and $\rho = \frac{Cov(Y_1,Y_2)}{\sigma_1\sigma_2}$.
   We can also directly verify that

$$E[Y_1] = E[\sigma_1 X_1 + \mu_1] = \sigma_1 E[X_1] + \mu_1 = 0 + \mu_1 = \mu_1$$

   and

$$E[Y_2] = E[\rho\sigma_2 X_1 + \sigma_2\sqrt{1-\rho^2}X_2 + \mu_2] = \rho\sigma_2 E[X_1] + \sigma_2\sqrt{1-\rho^2}E[X_2] + \mu_2 = 0 + 0 + \mu_2 = \mu_2$$

   since $E[X_1] = E[X_2] = 0$ since they are *i.i.d.* standard normal random variables.
   We can use the fact from lecture that

$$Var(Y_i) = a_{i1}^2 + a_{i2}^2$$

   to directly verify that

$$Var(Y_1) = \sigma_1^2 + 0^2 = \sigma_1^2$$

   and

$$Var(Y_2) = \rho^2\sigma_2^2 + \sigma_2^2(1-\rho^2) = \rho^2\sigma_2^2 + \sigma_2^2 - \rho^2\sigma_2^2 = \sigma_2^2$$

This allows us to use the fact that

$$Cov(Y_1, Y_2) = a_{11}a_{21} + a_{12}a_{22}$$

to compute that

$$\rho = \frac{\sigma_1\rho\sigma_2 + 0 \cdot \sigma_2\sqrt{1-\rho^2}}{\sigma_1\sigma_2} = \frac{\rho\sigma_1\sigma_2}{\sigma_2\sigma_2} = \rho$$

which completes the direct verification of the five parameters.

(b) We know from HW 11, Exercise 7, that if we let U_i = the result of the ith call to **Uniform(0,1)** and X_i = **InversePhi(U_i)**, then the cumulative distribution function of X_i is

$$F_{X_i}(x) = \begin{cases} 0 \text{ if } \Phi(x) = 0 \\ \Phi(x) \text{ if } 0 < \Phi(x) < 1 \\ 1 \text{ otherwise} \end{cases} = \Phi(x) \text{ for all } x \in \mathbb{R}$$

Thus, assuming the result of the ith call to **Uniform(0,1)** is independent from all other calls, we know X_1 and X_2 are *i.i.d.* standard normal random variables.

From part (a), we know that if we let

$$\begin{aligned} Y_1 &= \sigma_1 X_1 + \mu_1 \\ Y_2 &= \rho\sigma_2 X_1 + \sigma_2\sqrt{1-\rho^2}X_2 + \mu_2 \end{aligned}$$

then (Y_1, Y_2) is a bivariate normal random variable with parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$. This provides a very simple routine to sample from a bivariate normal random variable using two independent samples from **Uniform(0,1)**:

BivariateNormal(mu1, mu2, sigma1, sigma2, rho){
Let $u1 = Uniform(0, 1)$
Let $x1 = InversePhi(u1)$
Let $u2 = Uniform(0, 1)$
Let $x2 = InversePhi(u2)$
Let $y1 = sigma1 * x1 + mu1$
Let $y2 = rho * sigma2 * x2 + sigma2 * sqrt(1 - (rho * rho))x2 + mu2$
return $(y1, y2)$ }

3. Let f, g be real-valued functions. The *convolution* of f and g is the real-valued function $f * g$ defined by

$$(f * g)(a) = \int_{-\infty}^{\infty} f(a-y)g(y)\,dy.$$

(a) Suppose X, Y are jointly continuous, independent random variables. Show that

$$f_{X+Y} = f_X * f_Y.$$

That is, the PDF of the *independent* sum of continuous random variables is obtained by taking the convolution of their PDF's.

(Hint: as usual, your argument should use cumulative distribution functions and derivatives.)

(b) Suppose X, Y, Z are i.i.d. Uniform$(0, 1)$ random variables. Explicitly compute the PDF of $X+Y+Z$ using convolution.

(Aside: this is an *Irwin–Hall* distribution. While there are general formulas for the PDF of the sum of n i.i.d. Uniform$(0, 1)$ random variables, they are complicated and have many terms.)

(c) Suppose X, Y are i.i.d. Normal$(0, 1)$ random variables. Using convolution, verify that $X + Y \sim$ Normal$(0, \sqrt{2})$.

(Hint: recall the Gaussian integral, and how to complete the square.)

Solution.

(a) Since

$$f_{X+Y}(z) = \frac{d}{dz} F_{X+Y}(z)$$

we can compute f_{X+Y} by first computing F_{X+Y}. Note that

$$F_{X+Y}(z) = \mathbb{P}(X + Y \leq z) = \mathbb{P}(X \leq z - y, Y \leq y)$$

for all $-\infty < y < \infty$. Since X and Y are continuous random variables, we have

$$F_{X+Y}(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f_{X,Y}(x, y)dxdy$$

Since X and Y are independent, we have

$$F_{X+Y}(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f_X(x)f_Y(y)dxdy = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z-y} f_X(x)dx \right) f_Y(y)dy$$

Since

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(x)dx$$

we have

$$F_{X+Y}(z) = \int_{-\infty}^{\infty} F_X(z-y)f_Y(y)dy$$

Differentiating, we find that the density function of $X + Y$ is

$$\begin{aligned} f_{X+Y}(z) = & \quad \frac{d}{dz} \int_{-\infty}^{\infty} F_X(z-y)f_Y(y)dy = \int_{-\infty}^{\infty} \frac{d}{dz} F_X(z-y)f_Y(y)dy \\ = & \quad \int_{-\infty}^{\infty} f_X(z-y)f_y(y)dy = (f_X * f_Y)(z) \end{aligned}$$

Thus, we have

$$f_{X+Y}(z) = (f_X * f_Y)(z)$$

This completes the proof that, for jointly continuous, independent random variables, the PDF of the sum of the variables is obtained by taking the convolution of their PDF's.

(b) Since X, Y, and Z are *i.i.d.* Uniform(0,1), we know that

$$f_X(a) = f_Y(a) = f_Z(a) = \begin{cases} 1 \text{ if } 0 \leq x \leq 1 \\ 0 \text{ otherwise.} \end{cases}$$

Thus, we can apply part (a) to find

$$\begin{aligned} f_{X+Y}(a) = & \int_{-\infty}^{\infty} f_X(a-y) f_Y(y) dy = \int_0^1 f_X(a-y) f_y(y) dy \\ = & \begin{cases} \int_0^a dy \text{ if } 0 \leq a \leq 1 \\ \int_{a-1}^1 dy \text{ if } 1 \leq a \leq 2 \\ 0 \text{ otherwise.} \end{cases} \end{aligned}$$

We can directly compute that

$$\int_0^a dy = y\Big|_0^a = a - 0 = a$$

and

$$\int_{a-1}^1 dy = y\Big|_{a-1}^1 = 1 - (a-1) = 2 - a$$

so we know that

$$f_{X+Y}(a) = \begin{cases} a \text{ if } 0 \leq a \leq 1 \\ 2 - a \text{ if } 1 \leq a \leq 2 \\ 0 \text{ otherwise.} \end{cases}$$

We can know use part (a) again to compute f_{X+Y+Z} as the convolution of f_{X+Y} and f_Z. We have

$$f_{X+Y+Z}(b) = \int_{-\infty}^{\infty} f_{X+Y}(b-z) f_Z(z) dz = \int_0^1 f_{X+Y}(b-z) dz$$

If $0 \leq b \leq 1$, then $f_{X+Y}(b-z)$ is only nonzero for $0 \leq b - z \implies b \geq z$. Thus, we should only integrate from 0 to b for such b.
If $1 \leq b \leq 2$, then $f_{X+Y}(b-z)$ is nonzero for all of $z : 0 \to 1$. However, when $(b-z) \geq 1 \iff z \leq b-1$, we have

$$f_{X+Y}(b-z) = 2 - (b-z) = 2 - b + z$$

and when $(b-z) \leq 1 \iff z \geq b-1$, we have

$$f_{X+Y}(b-z) = b - z$$

Thus, we must split the integral into one from $z : 0 \to b-1$ and one from $z : b-1 \to 1$ for such b.
If $2 \leq b \leq 3$, then $f_{X+Y}(b-z)$ is only nonzero when $(b-z) \leq 2 \iff z \geq b-2$, so we should only integrate from $b-2$ to 1 for such b.
This combines to yield

$$f_{X+Y+Z}(b) = \begin{cases} \int_0^b (b-z) dz \text{ if } 0 \leq b \leq 1 \\ \int_0^{b-1} (2-b+z) dz + \int_{b-1}^1 (b-z) dz \text{ if } 1 \leq b \leq 2 \\ \int_{b-2}^1 (2-b+z) dz \text{ if } 2 \leq b \leq 3 \end{cases}$$

We can directly compute that

$$\int_0^b (b-z) dz = bz - \frac{z^2}{2}\Big|_0^b = b^2 - \frac{b^2}{2} = \frac{b^2}{2}$$

and

$$\begin{aligned}
&\int_0^{b-1}(2-b+z)dz+\int_{b-1}^{1}(b-z)dz\\
&=\left((2-b)z+\frac{z^2}{2}\right)\Big|_0^{b-1}+\left(bz-\frac{z^2}{2}\right)\Big|_{b-1}^{1}\\
&=(2-b)(b-1)+\frac{(b-1)^2}{2}+(b-\frac{1}{2})-(b(b-1)-\frac{(b-1)^2}{2})\\
&=(b-1)^2+2b-b^2-2+b+b-\frac{1}{2}-b^2+b\\
&=(b-1)^2+5b-2b^2-\frac{5}{2}\\
&=b^2-2b+1+5b-2b^2-\frac{5}{2}\\
&=-b^2+3b-\frac{3}{2}\\
&=\frac{1}{2}(-2b^2+6b-3)
\end{aligned}$$

and

$$\begin{aligned}
\int_{b-2}^{1}2-b+zdz= &\quad (2-b)z+\frac{z^2}{2}\Big|_{b-2}^{1}\\
= &\quad (2-b)+\frac{1}{2}-\left((2-b)(b-2)+\frac{(b-2)^2}{2}\right)\\
= &\quad \frac{5}{2}-b+\frac{(b-2)^2}{2}=\frac{1}{2}(5-2b+b^2-4b+4)=\frac{1}{2}(b^2-6b+9)\\
= &\quad \frac{1}{2}(b-3)^2
\end{aligned}$$

Thus, we can express $f_{X+Y+Z}(b)$ explicitly as

$$f_{X+Y+Z}(b)=\begin{cases}\frac{b^2}{2}\text{ if }0\le b\le 1\\ \frac{1}{2}(-2b^2+6b-3)\text{ if }1\le b\le 2\\ \frac{1}{2}(b-3)^2\text{ if }2\le b\le 3\\ 0\text{ otherwise.}\end{cases}$$

(c) Since X and Y are *i.i.d.* Normal(0,1) random variables, we know

$$f_X(a)=f_Y(a)=\frac{1}{\sqrt{2\pi}}exp(-\frac{a^2}{2})$$

so we can apply part (a) to find

$$\begin{aligned}
f_{X+Y}(a)= &\quad \int_{-\infty}^{\infty}f_X(y-a)f_Y(a)dy=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}exp(-\frac{(a-y)^2}{2})\frac{1}{\sqrt{2\pi}}exp(-\frac{y^2}{2})dy\\
= &\quad \frac{1}{2\pi}\int_{-\infty}^{\infty}exp(-\frac{(a-y)^2+y^2}{2})dy
\end{aligned}$$

Completing the square, we find

$$\frac{(a-y)^2+y^2}{2}=y^2-ay+\frac{a^2}{2}=(y-\frac{a}{2})^2+\frac{a^2}{4}$$

so we can apply the fact that $a^b a^c = a^{bc}$ to find

$$f_{X+Y}(a) = \frac{1}{2\pi}\int_{-\infty}^{\infty} exp(-(y-\frac{a}{2}) - \frac{a^2}{4})dy = \frac{exp(-\frac{a^2}{4})}{2\pi}\int_{-\infty}^{\infty} exp(-(y-\frac{a}{2})^2)dy$$

Applying the general Gaussian integral

$$\int_{-\infty}^{\infty} e^{-a(x+b)^2}dx = \sqrt{\frac{\pi}{a}}$$

we find

$$f_{X+Y}(a) = \frac{exp(-\frac{a^2}{4})}{2\pi}\sqrt{\pi} = \frac{1}{\sqrt{2\pi\sqrt{2}^2}}exp(-\frac{a^2}{2\sqrt{2}^2})$$

Note that this is just the PDF of a normal random variable with $\mu = 0$, $\sigma = \sqrt{2}$. This completes the proof that $X + Y \sim Normal(0, \sqrt{2})$.

4. Recall the fundamental simple linear regression model,

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

with three parameters β_0, β_1, σ where $\varepsilon \sim N(0, \sigma)$ is a normally distributed error term which is independent of X.

In 1903, Pearson and Lee collected data on the heights of 1078 pairs of fathers and sons. The father's heights had a sample mean of 67.7 and a sample standard deviation of 2.72, while the son's heights had a sample mean of 68.7 and a sample standard deviation of 2.82. The sample correlation between the father's heights and the son's heights was 0.50.

(a) If the father is 74 inches tall, what would you predict the son's height to be? (*Hint*: use the formulas from lecture to estimate β_0 and β_1. In particular, estimate $\rho(X, Y)$ using the sample correlation. Use these parameter estimates to compute $\hat{Y} = \beta_0 + \beta_1 X$.)

(b) In the situation in part (a), compute the z-score of the father's height and the z-score of the predicted value of the son's height. Does the result exhibit regression towards the mean?

Solution.
Let $X =$ the height of a given father and $Y =$ the height of his son.

(a) We know that

$$\beta_1 = \frac{\rho(X, Y)\sigma_Y}{\sigma_X}$$

so we can use our unbiased estimators for $\rho \approx 0.50$, $\sigma_X \approx 2.72$, and $\sigma_Y \approx 2.82$ to estimate that

$$\beta_1 \approx \frac{0.50 \cdot 2.82}{2.72} \approx 0.5184$$

We know that

$$\beta_0 = E[Y] - \beta_1 E[X]$$

so we can use our unbiased estimators for $E[Y] \approx 68.7$, $E[Y] \approx 67.7$, and $\beta_1 \approx 0.5184$ to estimate that

$$\beta_0 \approx 68.7 - 0.5184 \cdot 74 \approx 33.61$$

This allows us to calculate that

$$\hat{Y} = \beta_0 + \beta_1 X \approx 33.61 + 0.5184 \cdot 74 \approx 71.97$$

Thus, if the father is 74 inches tall, the son's height is predicted to be about 71.97 inches.

(b) We can quickly compute that the z-score of the father's height is

$$z_x = \frac{74 - 67.7}{2.72} \approx 2.316$$

while the z-score of the predicted value of the son's height is only

$$z_{\hat{y}} = \frac{71.97 - 68.7}{2.82} \approx 1.160$$

Thus, as the z-score of the predicted value of the son's height is about half the z-score of the father's height, and the sample correlation is 0.50, the result *does* exhibit regression towards the mean. The son's height is predicted to be about half as far from the mean as the father's.

5. (Ross, P7.80) The moment generating function of X is given by $M_X(t) = \exp(2e^t - 2)$ and that of Y by $M_Y(t) = \left(\frac{3}{4}e^t + \frac{1}{4}\right)^{10}$. If X and Y are independent, what are

   a $P(X + Y = 2)$?

   b $P(XY = 0)$?

   c $E[XY]$?

   Solution.
   Note: For any discrete random variable X, $M_X(t) = G_X(e^t)$, where $G_X(t)$ is the probability generating function of X.
   We know that a Poisson random variable P with parameter λ has probability generating function

$$G_P(t) = exp(\lambda(t-1))$$

   Thus, P has moment generating function

$$M_P(t) = exp(\lambda(e^t - 1))$$

   Since $M_X(t) = exp(2e^t - 2) = exp(2(e^t - 1)) = M_P(t)$ for $\lambda = 2$, and

$$M_A(t) = M_B(t) \implies f_A(t) = f_B(t)$$

   we know $X \sim Poisson(2)$. Thus, X has probability mass function

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

   Similarly, any Binomial random variable B with parameters n and p has probability generating function

$$G_B(t) = (pt + (1-p))^n$$

   so it has moment generating function

$$M_B(t) = (pe^t + (1-p))^n$$

   Since $M_Y(t) = (\frac{3}{4}e^t + \frac{1}{4})^{10} = M_B(t)$ for $n = 10$, $p = \frac{3}{4}$, we know $Y \sim Binomial(10, \frac{3}{4})$. Thus, Y has probability mass function

$$p_Y(k) = \binom{10}{k}(\frac{3}{4})^k(\frac{1}{4})^{10-k}$$

   (a) For $X + Y = 2$, we need either $X = 2$ and $Y = 0$, $X = 1$ and $Y = 1$, or $X = 0$ and $Y = 2$. These are all mutually disjoint events, which yields

$$\begin{aligned}
\mathbb{P}(X+Y=2) = &\quad \mathbb{P}(X=2,Y=0) + \mathbb{P}(X=1,Y=1) + \mathbb{P}(X=0,Y=2) \\
= &\quad \mathbb{P}(X=2)\mathbb{P}(Y=0) + \mathbb{P}(X=1)\mathbb{P}(Y=1) + \mathbb{P}(X=0)\mathbb{P}(Y=2) \\
= &\quad e^{-2}\frac{2^2}{2}\binom{10}{0}\frac{1}{4}^{10} + e^{-2}\frac{2^1}{1!}\binom{10}{1}\frac{3}{4}^1\frac{1}{4}^9 + e^{-2}\binom{10}{8}\frac{3}{4}^2\frac{1}{4}^8 \\
= &\quad 2e^{-2}\frac{1}{4}^{10} + 15e^{-2}\frac{1}{4}^9 + 45e^{-2}\frac{3}{4}^2\frac{1}{4}^8 \approx 0.00006 = 0.006\%
\end{aligned}$$

   (b) For $XY = 0$, we have to find $\mathbb{P}(X = 0 \cup Y = 0)$. We can apply the Principle of Inclusion Exclusion to find

$$\mathbb{P}(X = 0 \cup Y = 0) = \mathbb{P}(X = 0) + \mathbb{P}(Y = 0) - \mathbb{P}(X = 0, Y = 0)$$

   Now, using the fact that X and Y are independent, we find

$$\mathbb{P}(X = 0 \cup Y = 0) = e^{-2} + (\frac{1}{4})^{10} - e^{-2} \cdot (\frac{1}{4})^{10} = \frac{4^{10} + e^2 - 1}{4^{10}e^2} \approx 0.1353 = 13.53\%$$

(c) Since X and Y are independent, we know that

$$E[XY] = E[X]E[Y] = \frac{10 \cdot 3}{4} \cdot 2 = \frac{60}{4} = 15$$

so the expected value of XY is 15.

6. The *characteristic function* of a random variable X is the complex-valued function ϕ_X of a real variable t defined by $\phi_X(t) = E[e^{itX}]$. Explicitly,

$$\phi_X(t) = \begin{cases} \int_{-\infty}^{\infty} e^{itx} f(x)\,dx & X \text{ continuous} \\ \sum_x e^{itx} p(x) & X \text{ discrete,} \end{cases}$$

where f is the density of X or p is the mass function of X. (Here $e^{itx} = \cos(tx) + i\sin(tx)$ by Euler's formula.)

(a) Determine the characteristic function of a uniform continuous random variable on $[a, b]$.

(b) Determine the characteristic function of a uniform discrete random variable on $\{a, a+1, \ldots, b-1, b\}$. Your expression should be similar to your expression from (a).

(c) The characteristic function ϕ_X is the *Fourier transform* of the PDF of X. It is a general principle that under the Fourier transform, convolution corresponds to multiplication.
Let X, Y be i.i.d. jointly continuous random variables. Show directly that

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t).$$

Solution.

(a) Let $X \sim ContinuousUniform(a, b)$. Then X has PDF

$$f_X(x) = \begin{cases} \frac{1}{b-a} \text{ if } a \le x \le b \\ 0 \text{ otherwise.} \end{cases}$$

Applying the definition of the characteristic function of a continuous random variable, we find

$$\phi_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(x)dx = \int_a^b \frac{e^{itx}}{b-a}dx = \frac{1}{b-a}\left(\frac{e^{itx}}{it}\Bigg|_a^b\right) = \frac{e^{itb} - e^{ita}}{(b-a)it}$$

Note: This formula does not hold for $t = 0$, at which point it is undefined. However, we can clearly see that, when $t = 0$,

$$\phi_X(t) = \int_a^b \frac{1}{b-a}dx = \frac{b-a}{b-a} = 1$$

so we have

$$\phi_X(t) = \begin{cases} 1 \text{ if } t = 0 \\ \frac{e^{itb} - e^{ita}}{(b-a)it} \text{ otherwise.} \end{cases}$$

for a continuous uniform random variable.

(b) Now, let $X \sim DiscreteUniform(a, b)$. Then X has PMF

$$p_X(k) = \begin{cases} \frac{1}{b-a+1} \text{ if } k \in \{a, ..., b\} \\ 0 \text{ otherwise.} \end{cases}$$

Applying the definition of the characteristic function for a discrete random variable, we find

$$\begin{aligned} \phi_X(t) = & \sum_{k=a}^{b} \frac{e^{itk}}{b-a+1} = \frac{1}{b-a+1}\left(\sum_{k=0}^{b} e^{itk} - \sum_{k=0}^{a-1} e^{itk}\right) \\ = & \frac{1}{b-a+1}\left(\frac{1-e^{it(b+1}}{1-e^{it}} - \frac{1-e^{ita}}{1-e^{it}}\right) \\ = & \frac{e^{ita} - e^{it(b+1)}}{(b-a+1)(1-e^{it})} \end{aligned}$$

Note: This formula does not hold for $t = 0$, at which point it is undefined. However, we can clearly see that, when $t = 0$, we have

$$\phi_X(t) = \sum_a^b \frac{1}{b-a+1} = \frac{b-a+1}{b-a+1} = 1$$

Thus, we have

$$\phi_X(t) = \begin{cases} 1 \text{ if } t = 0 \\ \frac{e^{ita} - e^{it(b+1)}}{(b-a+1)(1-e^{it})} \text{ otherwise.} \end{cases}$$

for a discrete uniform random variable.

(c) We can use the fact that $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ since X and Y are independent continuous random variables. Applying the definition of the characteristic function, we find

$$\phi_{X+Y}(t) = E[e^{it(X+Y)}] = E[e^{itX+itY}] = E[e^{itX}e^{itY}]$$

If we let $f(X) = e^{itX}$ and $g(Y) = e^{itY}$, we clearly see that

$$\phi_{X+Y}(t) = E[f(x)g(Y)] = E[f(x)]E[g(Y)] = E[e^{itX}]E[e^{itY}] = \phi_X(t)\phi_Y(t)$$

This completes the proof that, for jointly continuous *i.i.d.* continuous random variables X and Y,

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$$

7. This problem formally introduces unbiased estimators for the parameters β_0, β_1, and σ^2 in simple linear regression. It has the virtue of using most of the ideas introduced in this course simultaneously and hence serves as something of a "capstone."

   We begin with random variables (X, Y) with some unknown joint distribution. We suppose that there are some parameters β_0, β_1, σ and a *true regression line* $\hat{y} = \beta_0 + \beta_1 x$ such that, for every fixed value $X = x$, the conditional distribution of the error in estimating Y using the true regression line is normal:

$$Y - \hat{y} \mid X = x \sim N(0, \sigma).$$

   In particular, $E[Y - (\beta_0 + \beta_1 X) \mid X = x] = 0$.

   We now take i.i.d. random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ with the same joint distribution as (X, Y). The following random variables will be used to estimate the parameters $\beta_0, \beta_1, \sigma^2$. Let

   - $\bar{X} = (X_1 + \cdots + X_n)/n$
   - $\bar{Y} = (Y_1 + \cdots + Y_n)/n$
   - $S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
   - $S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$
   - $r = \frac{1}{n-1} \sum_{i=1}^n \frac{X_i - \bar{X}}{S_X} \frac{Y_i - \bar{Y}}{S_Y}$
   - $b_1 = r S_Y / S_X$
   - $b_0 = \bar{Y} - b_1 \bar{X}$
   - $\hat{Y}_i = b_0 + b_1 X_i$
   - $s^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$

   Any experimentally obtained sample will consist of n pairs of data points $(x_1, y_1), \ldots, (x_n, y_n)$, and the random variables $\bar{X}, \bar{Y}, S_X^2, S_Y^2, r, b_1, b_0, \hat{Y}_i, s^2$ will all take on concrete values.

   You will show

$$E[b_0] = \beta_0, \qquad E[b_1] = \beta_1, \qquad E[s^2] = \sigma^2.$$

   (a) Show that

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

   (b) Write $b_1 \mid (X_j = x_j)$ for the conditional random variable $b_1 \mid (X_1 = x_1, \ldots, X_n = x_n)$. Show that

$$b_1 \mid (X_j = x_j) = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{(\sum_{i=1}^n x_i^2) - n\bar{x}^2},$$

   where $\bar{x} = \sum_{i=1}^n x_i / n$.

   (c) Show that

$$E[b_1 \mid (X_j = x_j)] = \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{(\sum_{i=1}^n x_i^2) - n\bar{x}^2}.$$

   (*Hint:* first show that $E[Y_i \mid (X_j = x_j)] = \beta_0 + \beta_1 x_i$.)

   (d) Show that

$$\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i) = \beta_1 \left(\left(\sum_{i=1}^n x_i^2 \right) - n\bar{x}^2 \right).$$

   Conclude that

$$E[b_1 \mid (X_j = x_j)] = \beta_1.$$

   (e) Conclude that

$$E[b_1] = \beta_1.$$

(f) Show that

$$E[b_0 \mid (X_j = x_j)] = \beta_0.$$

Conclude that

$$E[b_0] = \beta_0.$$

(g) Show that

$$b_1 \mid (X_j = x_j) \sim N\left(\beta_1, \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}\right).$$

(h) (Bonus.) Show that

$$\mathrm{Var}(b_0 \mid (X_j = x_j)) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}.$$

(*Hint:* $\mathrm{Cov}(\bar{Y}, b_1) = 0$.)

(i) (Bonus.) Show that

$$E[s^2] = \sigma^2.$$

In practice, (g) is used to test hypotheses like "X and Y are uncorrelated", i.e. $\beta_1 = 0$. Specifically, σ^2 is estimated by s^2 and b_1 is calculated for a given data set $(x_1, y_1), \ldots, (x_n, y_n)$. The z-score

$$z = \frac{b_1}{s/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

is then computed. If $\beta_1 = 0$, then this z-score would follow a standard normal $Z \sim N(0, 1)$. Now one computes the probability that a standard normal would be at least as far as z is from 0. This is called the *p-value* and here is $p = 2P(Z > |z|) = 2(1 - \Phi(|z|))$. Finally, if p is smaller than some threshold such as 0.05, the hypothesis that X and Y are uncorrelated is rejected and the data set yields fairly strong evidence that X and Y are correlated.

The father/son example produces an extraordinarily tiny p-value and therefore extremely strong evidence that father/son heights are correlated, in agreement with our intuition.

Solution.

(a) Note that

$$r = \frac{1}{(n-1) S_Y S_X} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

so

$$b_1 = r \frac{S_Y}{S_X} = \frac{1}{(n-1) S_X^2} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

Plugging in the given formula for S_X, we find

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1) \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

as required.

(b) If we are given $(X_j = x_j)$, we can rewrite b_1 as

$$b_1 | (X_j = x_j) = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i) - \bar{Y} \sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Since $\bar{X}$ is the raw average of $x_1, \ldots, x_n$, we know

$$\sum_{i=1}^{n}(x_i - \bar{x}) = (x_1 + \cdots + x_n) - (x_1 + \cdots + x_n) = 0$$

so we have

$$b_1|(X_j = x_j) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Note that

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n\bar{x}^2 = \sum_{i=1}^{n}x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2 = \sum_{i=1}^{n}(x_i^2) - n\bar{x}^2$$

This completes the proof that

$$b_1|(X_j = x_j) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i^2) - n\bar{x}^2}$$

(c) Note that, since everything besides the Y_is is a constant, we have

$$E[b|(X_j = x_j)] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E[Y_i]}{\sum_{i=1}^{n}(x_i^2) - n\bar{x}^2}$$

so we just need to compute $E[Y_i]$. Since $E[Y_i - (\beta_0 + \beta_1 x_i)] = 0$, we know

$$E[Y_i] = E[\beta_0 + \beta_1 x_i] = \beta_0 + \beta_1 x_i$$

since β_0, β_1, and x_i are constants given $(X_j = x_j)$. This yields

$$E[b|(X_j = x_j)] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^{n}(x_i^2) - n\bar{x}^2}$$

as required.

(d) We can split the sum to find

$$\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i) = \beta_0\sum_{i=1}^{n}(x_i - \bar{x}) + \beta_1\sum_{i=1}^{n}(x_i^2 - x_i\bar{x})$$

Since

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

we know

$$\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i) = \beta_1\sum_{i=1}^{n}(x_i^2 - x_i\bar{x}) = \beta_1\left(\sum_{i=1^n} x_i^2 - \bar{x}\sum_{i=1}^{n}x_i\right)$$

Since

$$\sum_{i=1}^{n}x_i = n\bar{x}$$

we know

$$\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i) = \beta_1\left((\sum_{i=1^n} x_i^2) - n\bar{x}^2\right)$$

as required. Plugging this into the equation from part (c), we immediately see that

$$E[b|(X_j = x_j)] = \frac{\beta_1\left((\sum_{i=1^n} x_i^2) - n\bar{x}^2\right)}{\sum_{i=1}^{n}(x_i^2) - n\bar{x}^2} = \beta_1$$

(e) Since

$$E[b_1|(X_j = x_j)] = \beta_1$$

for all X_j, we know that $E[b_1]$ does not depend on X_j. Thus, we can conclude that

$$E[b_1] = \beta_1$$

as required.

(f) Since

$$b_0 = \bar{Y} - b_1\bar{X}$$

we know

$$b_0|(X_j = x_j) = \bar{Y}|(X_j = x_j) - b_1|(X_j = x_j) \cdot \bar{x}$$

Note that

$$E[\bar{Y}|(X_j = x_j)] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i|(X_j = x_j)] = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) = \frac{n\beta_0}{n} + \beta_1 \sum_{i=1}^{n}\frac{x_i}{n} = \beta_0 + \beta_1\bar{x}$$

Since we already computed

$$E[b_1|(X_j = x_j)]$$

in part (c), we know

$$E[b_0|(X_j = x_j)] = \beta_0 + \beta_1\bar{x} - \beta 1\bar{x} = \beta_0$$

as required.

(g) From part (b), since everything involving x_i's and $\bar{X}$ is a constant given $(X_j = x_j)$, we can rewrite $b_1|(X_j = x_j)$ as

$$b_1|(X_j = x_j) = \sum_{i=1}^{n} c_i Y_i$$

Note: For a given $(X_j = x_j)$, we know that

$$Y_i - \hat{y} = Y_i - (\beta_0 + \beta_1 x_i) \sim Normal(0, \sigma)$$

Thus, $b_1|(X_j = x_j)$ is a weighted sum of *i.i.d.* Normal(0, σ) random variables, so $b_1|(X_j = X_j)$ is a normal random variable. Note that the weights

$$c_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

so we know

$$Var(b_1|(X_j = x_j)) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} Var(Y_i) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

This completes the proof that

$$b_1|(X_j = x_j) \sim Normal(\beta_1, \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}})$$

CSCI 270: Algorithms and Computing Theory

All assignments in this section were written by Shahriar Shamsian, Senior Lecturer of Computer Science, USC. Solutions to assignments 1 through 12 are provided.

Assignment 1

1.

Solve Kleinberg and Tardos, Chapter 1, Exercise 1:
Decide whether you think the following statement is true or false. If it is true, give a short explanation. If it is false, give a counterexample.
True or false? In every instance of the Stable Matching Problem, there is a stable matching containing a pair (m, w) such that m is ranked first on the preference list of w and w is ranked first on the preference list of m.

Claim: The statement is false.
Counterexample: Consider the following preference lists with $n = 2$ and $a : b > c$ indicating that a ranks b higher than c.

$$M_1 : W_1 > W_2 \qquad W_1 : M_2 > M_1$$
$$M_2 : W_2 > W_1 \qquad W_2 : M_1 > W_1$$

Note: There is no pair (M_i, W_j) such that M_i is ranked first on W_j's preference list and W_j is ranked first on M_i's preference list.
No stable matching can contain such a pair (M_i, W_j), as no such pair exists.
Therefore, the statement is false.

2.

Determine whether the following statement is true or false. If it is true, give an example. If it is false, give a short explanation. (5pts)
For some $n \geq 2$, there exists a set of preferences for n men and n women such that in the stable matching returned by the G-S algorithm when men are proposing, every woman is matched with their most preferred man, even though that man does not prefer that woman the most.

Claim: The statement is true.
Example: Consider the following preference lists with $n = 3$ and $a : b > c$ indicating a ranks b higher than c.

$$M_1 : W_1 > W_3 > W_2 \qquad W_1 : M_3 > M_1 > M_2$$
$$M_2 : W_1 > W_2 > W_3 \qquad W_2 : M_2 > M_3 > M_1$$
$$M_3 : W_2 > W_1 > W_3 \qquad W_3 : M_1 > M_2 > M_3$$

Now, let's trace the Gale-Shapley Algorithm with this scenario:
First, M_1 proposes to W_1, and they get engaged.
Next, M_2 proposes to W_1, but he gets rejected.
Next, M_2 proposes to W_2, and they get engaged.
Next, M_3 proposes to W_2, and he gets rejected.
Next, M_3 proposes to W_1, and they get engaged, breaking up M_1's engagement.
Finally, M_1 proposes to W_3, and they get engaged

Now, all men and all women are in exactly one engagement, so all active engagements are finalized into marriages.
In the end, (M_1, W_3), (M_2, W_2), and (M_3, W_1) are the three final marriages returned by the Gale Shapley algorithm. In each marriage, the woman is matched with her most preferred man, while the man is not matched with his most preferred woman.
Therefore, the statement is true.

3.

Solve Kleinberg and Tardos, Chapter 1, Exercise 4. (15 pts)
Show that there is always a stable assignment of students to hospitals, and give an algorithm to find one.

Note 1: We will use the definition of a stable matching provided by the problem description.
Note 2: Based on the problem description in the textbook, we will assume that all students rank all hospitals and all hospitals rank all students.

Now, let's provide some other necessary definitions:
Define $H :=$ the set of all hospitals.
Define $S :=$ the set of all students.
Define $f(h) :=$ # of roles filled at hospital $h \in H$.
Define $n(h) :=$ # of total roles at hospital $h \in H$.
Define $H_s :=$ the ordered ranking of hospitals by preference of student $s \in S$.
Define $S_h :=$ the ordered ranking of students by preference of hospital $h \in H$.
Define $(s, h) :=$ student s is assigned to hospital h.

Now, we will present an algorithm that always returns a stable matching of students and hospitals, that is, it returns a matching such that:
i) all roles at all hospitals are filled
ii) no assignment (s, h) exists such that hospital h prefers unassigned student s' to student s.
iii) No two assignments $(s, h), (s', h')$ exist such that hospital h prefers student s' to student s and student s' prefers hospital h to hospital h'

Our algorithm works as follows:

while $\exists$ unassigned $s \in S$ that and has NOT been unassigned from all $h \in H$
  assign such a student $s \in S$ to most preferred hospital $h \in H_s$
  if$(f(h) > n(h))$
    unassign student $s' \in S$ assigned to h that is least preferred on S_h
    remove h from $H_{s'}$
  endif
endwhile

Proof of Correctness: First, we need to make a helpful observation:
Observation 1: Since we check if a hospital $h \in H$ has too many students assigned to it every time we assign a new student, and we only remove a student if $f(h) > n(h)$, unassignment will never cause a filled role to open up. Therefore, once a role is filled, our algorithm will never open it up.

Now, we will show our algorithm satisfies all three necessary properties:
i) all roles at all hospitals are filled:
Assume to the contrary that our algorithm terminates and there exists an unfilled role at a hospital, that is, $\exists\ h \in H$ such that $f(h) < n(h)$.
There is a surplus of students, which implies that must exist either an unassigned student OR a hospital

with more students than total roles, or both.
Since the algorithm terminates only after all unassigned students have been unassigned from all $h \in H$, the existence of an unassigned student $s \in S$ implies that the algorithm assigned and then unassigned s to every hospital.
Unassignment only happens when all roles at a hospital are filled.
Therefore, the existence of an unassigned student implies that all roles at all the hospitals were at one point filled.
But since **Observation 1** tells us that our algorithm will never open up a role after filling it, this directly contradicts the existence of an open role at the termination of our algorithm.
On the other hand, the existence of a hospital with more students than total roles at the termination of the algorithm implies that a student was assigned to a full hospital without a student immediately being removed from that hospital.
However, assigning a student to a full hospital $h \in H$ will always cause $f(h) > n(h)$, which immediately causes our algorithm to unassign a student from h, another contradiction.
Therefore, by contradiction, our algorithm fills all hospital roles.

ii) no assignment (s, h) exists such that hospital h prefers unassigned student s' to student s:
Assume to the contrary that $\exists(s, h)$ such that hospital h prefers unassigned student s' to student s after the algorithm terminates.
Since the algorithm terminates after all students are either assigned or have been unassigned from all $h \in H$, and s' is unassigned, we know s' was unassigned from h, either before or after s was assigned to h.
If s' was unassigned from h after s was assigned to h, then the algorithm would have unassigned s instead of s' since it unassigns the least preferred student among those assigned to h, and h prefers s' over s.
If s' was unassigned from h before s was assigned to h, then there must have been $n(h)$ students assigned to h that were preferred over s'.
Observation 1 guarantees that all $n(h)$ of those roles will never open up, so there will still be $n(h)$ students assigned to h when s is assigned.
Hospital h will only unassign a student in favor of a more preferred student.
Therefore, since all of the initial $n(h)$ students assigned to h were preferred over s', all of the $n(h)$ students assigned to h when s is assigned are also preferred over s'.
Since h prefers s' to s, h also prefers all of the $n(h)$ students assigned to it over s.
Therefore, when s is assigned to h, it becomes the least preferred student among $n(h) + 1$ students assigned to h, which immediately causes s to be unassigned from h and h to be removed from H_s.
Student s can only be assigned to hospitals in H_s, so this implies s will never be assigned to h again.
However, this directly contradicts our assumption that (s, h) exists. By contradiction, our algorithm satisfies property (ii).

iii) No two assignments $(s, h), (s', h')$ exist such that hospital h prefers student s' to student s and student s' prefers hospital h to hospital h':
Assume to the contrary that two assignments $(s, h), (s', h')$ exist such that hospital h prefers student s' to student s and student s' prefers hospital h to hospital h'.
Student s' was either unassigned from hospital h at one point, or student s' was never assigned to h.
If student s' was unassigned from hospital h, then there must be $n(h)$ students assigned to hospital h that hospital h prefers over s'.
Since s is assigned to h, and we already proved that no hospital can have more students assigned than total roles (after the algorithm terminates), we know that s is one of these $n(h)$ students that h prefers over s'.
However, this directly contradicts our assumption that h prefers s' over s.
On the other hand, since students are assigned to their most preferred hospital from which they have not been unassigned, and we know s' is assigned to h', if s' was never assigned to h, then s' must prefer h' to h.
However, this directly contradicts our assumption that s' prefers h over h'.
Thus, by contradiction, our algorithm satisfies property (iii).

Therefore, we have proven that our algorithm always returns a stable matching of hospitals and students,

which implies that there always exists a stable matching of hospitals and students.

Time Complexity Analysis
Suppose there are n students, m hospitals, and $k < n$ total hospital roles. We assume that $k < n$ because the problem description identifies a surplus of students

Worst Case:
All of the $n-k$ unassigned students are assigned to and then unassigned from each hospital. Therefore, each of these unassigned students causes m iterations. In the worst case, each of the k assigned students ends up assigned to their least preferred hospital. Since students are assigned to hospitals in order of preference, this means each of these k assigned students was also assigned to each of the m hospitals at some point by the algorithm. Thus, all of the n students directly cause m iterations, for a total of $n \cdot m$ iterations. Since all functionality inside each iteration can operate in constant time with proper implementation, this means our algorithm has $O(n * m)$ worst case time complexity.

Best Case: In the best case, one hospital, h_i, will have $n(h_i) + 1$ students who rank h_i as their favorite hospital. For all $h \neq h_i$, there will be exactly $n(h)$ students who rank h as their favorite hospital. In this case, there is just one unassigned student. All $n-1$ assigned students will just be assigned to their favorite hospital, only taking one iteration each for a total of $n-1$ iterations. The one unassigned student will be assigned to and then unassigned from all $h \in H$, taking a total of m iterations. Thus, the total number of iterations in the best case is $n-1+m$. Thus, the time complexity of our algorithm in the best case is $O(n+m)$.

4.

Solve Kleinberg and Tardos, Chapter 1, Exercise 8. (10pts)
Resolve this question by doing one of the following two things:
(a) Give a proof that, for any set of preference lists, switching the order of a pair on the list cannot improve a woman's partner in the Gale-Shapley algorithm; or
(b) Give an example of a set of preference lists for which there is a switch that would improve the partner of a woman who switched preferences

I will complete option (b).
Consider the following truthful preference lists with $n = 3$:

$$M_1 : W_1 > W_2 > W_3 \qquad W_1 : M_3 > M_2 > M_1$$
$$M_2 : W_1 > W_3 > W_2 \qquad W_2 : M_2 > M_1 > M_3$$
$$M_3 : W_3 > W_1 > W_2 \qquad W_3 : M_2 > M_3 > M_1$$

Now, let's trace the Gale-Shapley Algorithm with this scenario:
First, M_1 proposes to W_1, and they get engaged.
Next, M_2 proposes to W_1, and they get engaged, breaking up M_1's engagement.
Next, M_1 proposes to W_2, and they get engaged.
Next, M_3 proposes to W_3, and they get engaged.
Now, all men and all women belong to exactly one engagement, so all engagements are finalized into marriages.
In the end (M_2, W_1), (M_1, W_2), and (M_3, W_3) are the three final marriages. By telling the truth, W_1 ends up married to her second choice man, M_2.

Let's now examine what happens if W_1 lies and says that she prefers M_1 to M_2. This results in the

untruthful preference lists:

$$M_1 : W_1 > W_2 > W_3 \qquad W_1 : M_3 > M_1 > M_2$$
$$M_2 : W_1 > W_3 > W_2 \qquad W_2 : M_2 > M_1 > M_3$$
$$M_3 : W_3 > W_1 > W_2 \qquad W_3 : M_2 > M_3 > M_1$$

Now, let's trace the Gale-Shapley Algorithm again:
First, M_1 proposes to W_1, and they get engaged.
Next, M_2 proposes to W_1, and he gets rejected.
Next, M_2 proposes to W_3, and they get engaged.
Next, M_3 proposes to W_3, and he gets rejected.
Next, M_3 proposes to W_1, and they get engaged, breaking up M_1's engagement.
Next, M_1 proposes to W_2, and they get engaged.
Now, all men and all women belong to exactly one engagement, so all engagements are finalized into marriages.
In the end, (M_3, W_1), (M_1, W_2), and (M_2, W_3) are the three final marriages. By lying, W_1 ends up married to her first choice man, M_3.

Therefore, with these lists of preferences, W_1 can directly improve her final partner by lying about one of her preferences, which concludes the example.

Assignment 2

1.

Arrange these functions under the O notation using only = (equivalent) or $\subseteq$ (strict subset of):
a) $2^{log(n)}$
b) 2^{3n}
c) $n^{nlog(n)}$
d) $log(n)$
e) $nlog(n^2)$
f) n^{n^2}
g) $log(log(n^n))$
All logs are base 2. (10 pts)

Solution. First we will manipulate some of the functions to make the arrangement more obvious:

a) $2^{log(n)} = n$ (polynomial)
b) $2^{3n} = (2^3)^n = 8^n$ (exponential)
c) $n^{nlog(n)} > n^n$ for all $n > 2$, and $n^n > 8^n$ for all $n > 8$, so $n^{nlog(n)} > 8^n$ for all $n > 8$ (exponential)
d) $log(n)$ (logarithmic)
e) $nlog(n^2) = 2nlog(n) > n$ for all $n > 2$ (polynomial)
f) $n > log(n)$ for all $n > 1 \implies n^2 > nlog(n) \implies n^{n^2} > n^{nlog(n)}$ for all $n > 1$ (exponential)
g) $log(log(n^n)) = log(nlog(n))$, and $nlog(n) > n$ for all $n > 2$, so $log(log(n^n)) > log(n)$ for all $n > 2$ (logarithmic)

Note: if $l(n)$ is a logarithmic function, $p(n)$ is a polynomial function, and $e(n)$ is an exponential function, then $O(l(n)) \subseteq O(p(n)) \subseteq O(e(n))$.
Combining this fact with our previous manipulations, we arrive at the following arrangement:

$$\underbrace{O(log(n))}_{(d)} \subseteq \underbrace{O(log(log(n^n))}_{(g)} \subseteq \underbrace{O(2^{log(n)})}_{(a)} \subseteq \underbrace{O(nlog(n^2))}_{(e)} \subseteq \underbrace{O(2^{3n})}_{(b)} \subseteq \underbrace{O(n^{nlog(n)})}_{(c)} \subseteq \underbrace{O(n^{n^2})}_{(f)}$$

2.

Given functions f_1, f_2, g_1, g_2 such that $f_1(n) = O(g_1(n))$ and $f_2(n) = O(g_2(n))$. For each of the following statements, decide whether it is true or false and briefly explain why. (12 pts)

a) $f_1(n)/f_2(n) = O(g_1(n)/g_2(n))$
b) $f_1(n) + f_2(n) = O(max(g_1(n), g_2(n)))$
c) $f_1(n)^2 = O(g_1(n)^2)$
d) $log_2(f_1(n)) = O(log_2(g_1(n)))$

Solution.

(a) **False.** Let $f_1(n) = n^2$, $f_2(n) = n$, $g_1(n) = n^3$, and $g_2(n) = n^3$.
Then $f_1(n) = O(g_1(n))$ and $f_2(n) = O(g_2(n))$, as required.
But $f_1(n)/f_2(n) = n^2/n = n$, and $g_1(n)/g_2(n) = n^3/n^3 = 1$.
$n \neq O(1)$, so this counterexample disproves statement (a).

(b) **True.** Since $f_1(n) = O(g_1(n))$ and $f_2(n) = O(g_2(n))$, we know that:

$$\exists n_1, n_2, c_1, c_2 > 0 \text{ such that } f_1(n) \leq c_1 g_1(n) \text{ for all } n \geq n_1 \text{ and} f_2(n) \leq c_2 g_2(n) \text{ for all } n \geq n_2$$

Therefore,

$$f_1(n) + f_2(n) \leq c_1 g_1(n) + c_2 g_2(n) \text{ for all } n \geq max(n_1, n_2) = n_3$$

Also,

$$c_1 g_1(n) + c_2 g_2(n) \leq c_1 max(g_1(n), g_2(n)) + c_2 max(g_1(n), g_2(n)) = (c_1 + c_2) max(g_1(n), g_2(n))$$

Thus, if we let $c_3 = c_1 + c_2$, we find:

$$f_1(n) + f_2(n) \leq c_3 max(g_1(n), g_2(n)) \text{ for all } n \geq n_3 \implies f_1(n) + f_2(n) = O(max(g_1(n), g_2(n))$$

which completes the proof of statement (b).

(c) **True.** Since $f_1(n) = O(g_1(n))$, we know

$$\exists n_1, c_1 > 0 \text{ such that } f_1(n) \leq c_1 g_1(n) \text{ for all } n \geq n_1$$

Therefore,

$$f_1(n)^2 \leq c_1^2 g_1(n)^2 \text{ for all } n \geq n_1$$

Let $c_2 = c_1^2$, and we find:

$$f_1(n)^2 \leq c_2 g_1(n)^2 \text{ for all } n \geq n_1 \implies f_1(n)^2 = O(g_1(n)^2))$$

which completes the proof of statement (c).

(d) **False.** Let $f_1(n) = 8$ and $g_1(n) = 1$. Then $f_1(n) = O(g_1(n))$, as required.
However, $log_2(f_1(n)) = log_2(8) = 3$, and $log_2(g_1(n)) = log_2(1) = 0$.
$3 \neq O(0)$, so this counteraxample disproves statement (d).

3.

Given an undirected graph G with n nodes and m edges, design an $O(m + n)$ algorithm to detect whether G contains a cycle. Your algorithm should output a cycle if there is one. (12 pts)

Solution. We will implement a modification of recursive Depth-First Search. We will use a helper function to implement this modification, and our primary function will just call this function until all the nodes are explored or a cycle is found. We will use an adjacency list representation of edges to traverse the edges incident to a given vertex v in $O(deg(v))$ time. We will also keep track of the parent of each node in an array to print the cycle (if found) in linear time relative to the length of the cycle.

```
boolean SearchComp( vertex current, vertex parent )
    if explored[current] is true
        return false
    endIf
    set explored[current] to true
    set parent[current] to parent
    for each edge (current, v) incident to current
        if explored[v] is false
            return SearchComp(v, current)
        endIf
        else
            if parent[current] != v
                vertex temp = current
                while parent[temp] != null
                    print temp
                    temp = parent[temp]
                endWhile
                return true
            endIf
        endElse
    endFor
    return false
endSearchComp

boolean hasCycle( G(V,E) )
    set parent[] to null
    set explored[] to null
    for int v: 1 → N
        pick vertex v
        if searchComp(v, null)
            return true
        endIf
    endFor
    return false
endHasCycle
```

Time Complexity: DFS runs in $O(m + n)$ time over a connected component with n vertices and m edges. For each call to our modified DFS (searchComp), we only ever stop the recursion early (if we find a cycle), we never cause more recursion to take place than in normal DFS. Therefore, searchComp will only run slower than $O(m+n)$ if the process underwent upon finding a cycle takes more than $O(m+n)$. However, printing a node takes constant time, and the while loop could only possibly run through every vertex in the connected component, for a total of $O(n)$ runtime. Therefore, the worst case runtime of searchComp, just

like recursive DFS, is still $O(m+n+n) = O(m+2n) = O(m+n)$.

For hasCycle, all of the intializations take a total of $O(N+N) = O(2N) = O(N)$ runtime. If there exists a cycle, searchComp will return true, so hasCycle will terminate early. Thus, we must consider what happens if there exists no cycle. In this case, we must iterate through the entire for loop. In the absence of a cycle, searchComp is just DFS, which will explore all nodes in a connected component. Therefore, searchComp will execute in constant time if the current vertex was a part of a previously explored connected component. Thus, for each connected component, the searchComp calls to the vertices in that component will take a total of $O(m+n)$ time. If there are k connected components, with $n_1, ..., n_k$ vertices and $m_1, ..., m_k$ edges respectively, then the sum of calling searchComp on each of these components will take $O((n_1+m_1) + ... + (n_k+m_k)) = O(N+M)$ time. Thus, in the worst case, hasCycle will terminate in $O(M+N)$ time, as required.

Note: We are redefining $N :=$ total number of vertices in graph and $M :=$ total number of edges in the graph to let n and m denote similar quantities for individual connected components.

4.

Solve Kleinberg and Tardos, Chapter 2, Exercise 6

Solution.
(a) We choose $f(n) = n^3$. The outermost for loop has exactly n iterations. The innermost for loop has a maximum of $(n-1)$ iterations. Inside each iteration, a maximum of $(n+1)$ steps are done (n addition steps + 1 storage step). Thus, there will never be more than $n(n-1)(n+1) = n(n^2-1) = n^3 - n = O(n^3)$ steps in the algorithm. Thus, the algorithm is upper bounded by $O(f(n)) = O(n^3)$.

(b) Note: There are $\frac{n}{4}$ values of i for which $i \leq \frac{n}{4}$. For each such value of i, there will be $\frac{n}{4}$ values of j for which $j \geq \frac{3n}{4}$. In each iteration with such a combination of i and j, there is at least $\frac{3n}{4} - \frac{n}{4} = \frac{2n}{4} = \frac{n}{2}$ work done adding up the entries of A[]. Thus, there is at least $\frac{n}{4}\frac{n}{4}\frac{n}{2} = \frac{n^3}{32}$ work done by the algorithm. Thus, the algorithm is also $\Omega(f(n)) = \Omega(n^3)$. This combines with part (a) to show the algorithm is $\Theta(f(n)) = \Theta(n^3)$.

(c) The modified algorithm works as follows:

for $i : 1 \to n$
    let $B(i,i) = A(i)$
    for $j : i+1 \to n$
        let B(i,j) = B(i,j-1) + A(j)
    endFor
endFor

Proof of Correctness: We only care about values of B(i,j) where $j > i$, so we just need to show that $B(i, j-1) + A(j) = A(i) + A(i+1) + \cdots + A(j-1) + A(j)$.
We will do so by induction on j.
Base Case: $j = i+1$, $B(i,j) = B(i,j-1) + A(j) = B(i,i) + A(j) = A(i) + A(j)$ as expected.
Inductive Hypothesis: Assume that $B(i,j-1) + A(j) = A(i) + A(i+1) + \cdots + A(j-1) + A(j)$ for all $i < j = k < n$.
Inductive Step: We want to show that $B(i,k+1) = A(i) + A(i+1) + \cdots + A(k) + A(k+1)$.
Our algorithm sets $B(i,k+1) = B(i,k) + A(k+1)$.
By our Inductive Hypothesis, we know $B(i,k) = A(i) + A(i+1) + \cdots + A(k-1) + A(k)$.
Therefore, our algorithm sets

$$B(i,k+1) = A(i) + A(i+1) + \cdots + A(k-1) + A(k) + A(k+1)$$

as required.

Time Complexity: Each iteration takes constant time. The inner for loop runs $(n - i)$ times for a given $1 \leq i \leq n$. Since i ranges from 1 to n, this means the number of iterations of the inner loop ranges from 0 to $n - 1$. Thus, the total number of iterations is

$$\sum_{k=0}^{n-1} k = \frac{n(n-1)}{2} = \frac{n^2}{2} - \frac{n}{2}$$

Since each iteration takes constant time, this means the total runtime of the algorithm is $O(n^2)$, which is faster than the previous algorithm.

5.

What Mathematicians often keep track of a statistic called their Erdos Number, after the great 20th century mathematician. Paul Erdos himself has a number of zero. Anyone who wrote a mathematical paper with him has a number of one, anyone who wrote a paper with someone who wrote a paper with him has a number of two, and so forth and so on. Supposing that we have a database of all mathematical papers ever written along with their authors: (6 pts)
a. Explain how to represent this data as a graph
b. Explain how we would compute the Erdos number for a particular researcher
c. Explain how we would determine all researcher with Erdos number at most two.

Solution.
(a) We could represent the data as a graph by storing each individual researcher as a distinct vertex in the graph. Then, we could create undirected edges between each pair of researchers that wrote a paper together.

(b) To compute the Erdos number for a particular researcher, we could just run a Breadth-First Search algorithm with the particular researcher as the start vertex and Erdos as the end vertex. This would trace the shortest path between the particular researcher and Erdos, the length of which would be that researcher's Erdos number.

(c) To determine all researchers with Erdos number at most two, we could run a modified breadth-first search algorithm with Erdos as the start vertex. The modification would entail stopping the search algorithm after completing level 2 of the BFS tree. Every researcher marked as explored by this algorithm would be guaranteed to have an Erdos number at most two. Even better, their specific Erdos number could be determined by just looking at what level of the BFS tree they are in.

6.

Given a DAG, give a linear-time algorithm to determine if there is a simple path that visits all vertices. (8 pts)

Solution.
We slightly modify the explained $O(m + n)$ topological ordering algorithm from Chapter 3 of the textbook. Our algorithm works as follows:

bool path(V,E)
    count = 0
    stack S = null
    for each node $v \in V$
        if indegree[v] == 0

```
            increment count by 1
            add v to S
        endIf
    endFor
    if count > 1
        return false
    endIf
    while S != null
        pop top node s ∈ S
        dependentcount = 0
        for all edges (u, s) outgoing from s
            decrement indegree[u] by 1
            if indegree[u] == 0
                add u to S
                increment dependentcount by 1
            endIf
        endFor
        if dependentcount > 1
            return false
        endIf
    endWhile
    return true
endPath
```

Time Complexity: The topological ordering algorithm from Chapter 3 runs in $O(m + n)$ time. This algorithm only makes 2 significant modifications to the runtime of the topological ordering algorithm. First, it adds a $O(n)$ for loop that checks for vertices with in-degree 0. Next, it stops the main $O(m + n)$ loop early if there are ever two vertices with no incoming edges. Thus, our algorithm's main loop can only ever run quicker than the main loop in the topological ordering algorithm, so its runtime is also $O(m + n)$. This means the total runtime of our algorithm is $O(m + n + n) = O(2n + m) = O(m + n)$, so it runs in linear time as required.

Assignment 3

1.

Suppose you want to drive from USC to Santa Monica. Your gas tank, when full, holds enough gas to go p miles. Suppose there are n gas stations along the route at distances $d_1 \leq d_2 \leq \cdots \leq d_n$ from USC. Assume that the distance between any neighboring gas stations, and the distance between USC and the first gas station, as well as the distance between the last gas station and Santa Monica, are all at most p miles. Assume you start from USC with the tank full. Your goal is to make as few gas stops as possible along the way. Give the most efficient algorithm to determine which gas stations you should stop at and prove that your algorithm yields an optimal solution (i.e., the minimum number of gas stops). Give the time complexity of your algorithm as a function of n. (15 points)

Solution. Our strategy is to only stop at the furthest away gas station which we can reach. At every stop, we will completely fill our tank. This ensures we go as far as possible on each gas tank before filling up, which should minimize the total number of stops needed. We can implement the algorithm as follows:

```
vector< int > leastStops (vector< int > d)
    let d_{n+1} = the distance from USC to Santa Monica
    add d_{n+1} to d
```

```
    sort d in ascending order
    set stops = an empty vector
    set traveled = 0
    set burnt = 0
    set gas = p
    for i : 1 → n
        decrement gas by d_i − (burnt + traveled)
        burnt = d_i − traveled
        if (gas − (d_{i+1} − d_i) < 0)
            set gas = p
            set burnt = 0
            set traveled = d_i
            add d_i to stops
        endIf
    endFor
    return stops
endLeastStops
```

Proof of Correctness:
First, we will show that our algorithm implements the described Greedy approach. Indeed, our gas tank starts at full with enough gas to go p miles. Traveled tracks the distance from USC to the gas station we last stopped at. Burnt tracks the number of miles we have driven since the last stop. Gas tracks the number of miles left in our tank at each gas station. We only ever refill the tank if we do not have enough gas to reach the next station (i.e. the number of additional miles we need to drive is greater than gas), at which we add that gas station to the list of stops. Thus, our algorithm does implement the Greedy approach of only stopping at the furthest gas station we can reach with our current gas.

Now, we will show that our algorithm always produces an optimal solution.

First, we must show that our algorithm always chooses sufficient stops for us to reach Santa Monica, as the optimal solution will never result in us running out of gas. It suffices to show that our algorithm always chooses stops such that our car has enough gas to get to the station i. We can do this via induction on i.
Base Case:
$i = 1$. Our algorithm initializes our gas tank to be full (p miles of range). The problem description guarantees that the distance between the first stop and USC is less than or equal to p miles. Therefore, we will have enough gas to get to the first station (station $i = 1$).
Inductive Hypothesis:
Assume that our algorithm always chooses stops such that we have enough gas to get to the station i for all $1 \leq i \leq k < n + 1$
Inductive Step:
Consider $i = k + 1$. By the *Inductive Hypothesis*, we know our algorithm chose stops such that we have enough gas to get to station k. Once we arrive at station station k:
1) if we already have enough gas to get to station $k + 1$, then the inductive step is trivially true
2) if we don't have enough gas to get to station $k + 1$, then we fill up our gas tank with p miles of range.
No two gas stations are more than p miles apart, so we now have enough gas to get to station $k + 1$.
By induction, our algorithm always chooses stops such that we have enough gas to get to station i for all $1 \leq i \leq n + 1$, where station $n + 1$ is Santa Monica. Therefore, our algorithm will always choose sufficient stops for us to get to Santa Monica without running out of gas.

Next, we must show that our solution always stays ahead of the optimal solution.
Consider $S :=$ the set of stops returned by our **leastStops()** algorithm, and
$O :=$ the set of stops in the optimal solution.
We want to show that, for each $s_i \in S$, $s_i \geq o_i$. We can do this via induction, where the size of S is $|S| = k$.

Base Case:
$i = 1$. Our algorithm's first stop is the furthest gas station we can reach from USC on a full tank of gas. Therefore, s_1 must be at least as far from USC as o_1, as it is impossible to reach any further stops without stopping earlier.
Inductive Hypothesis:
Assume that $s_i \geq o_i$ for all $1 \leq i \leq j < k$.
Inductive Step: Consider $i = j + 1$. By our *Inductive Hypothesis*, we know that $s_j \geq o_j$. Once we stop at s_j, we know our algorithm doesn't stop again until it reaches the furthest station it can reach without running out of gas. Thus s_{j+1} is the furthest station we can reach from s_j on a full tank of gas. Since $s_j \geq o_j$, we know that o_j cannot reach a station further than s_{j+1}. Thus, since the optimal solution never runs out of gas, we know $s_{j+1} \geq o_{j+1}$.
By induction, we know that $s_i \geq o_i$ for all $1 \leq i \leq k$.

Now, we just need to show that $|S| = |O$.
Assume to the contrary that $|S| \neq |O|$. Since O is optimal, this directly implies $|O| < |S|$. Let $|O| = m$. By the previous proof, we know $s_m \geq o_m$. Since $|O| < |S|$, we know $\exists s_{m+1} \in S$. Since **leastStops()** only adds a stop when we cannot get to the next station, we know we cannot get to Santa Monica from s_m on a full tank of gas. However, since o_m is the last stop in the O, and O must get us to Santa Monica, this implies we can get to Santa Monica from o_m on a full tank of gas. However, since $o_m \leq s_m$, this implies we can also get to Santa Monica from s_m on a full tank of gas. This is a contradiction, which proves $|O| = |S|$.

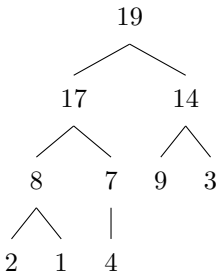
Thus, we have shown that $|O| = |S|$, which concludes the proof that **leastStops()** always yields an optimal solution.

Time Complexity Analysis:
The time complexity of **leastStops()** directly depends on the format of the input data. If the input data is already sorted in ascending order, then we could simply append d_{n+1} to the end of that list in $O(1)$ time to produce a sorted list of all $n + 1$ distances from USC in ascending order. If the data is not already sorted in ascending order, then we have to manually sort the data after appending d_{n+1}, which takes $O(nlog(n))$ time.
The runtime of the rest of the algorithm does not depend on this detail. Each iteration of the for loop only involves a constant number of constant time steps. Thus, each iteration of the for loop takes $O(1)$ time. The for loop always iterates n times, so it always takes $O(n)$ time in total.
This results in two distinct runtimes depending on the format of the input data.
1 - Sorted Input Data: $O(1) + O(n) = O(n)$ total runtime.
2 - Unsorted Input Data: $O(nlog(n)) + O(n) = O(nlog(n))$ total runtime.

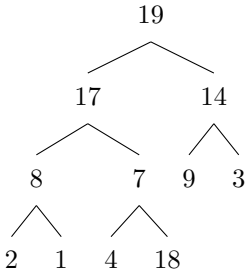
2.

The array A holds a max-heap. What will be the order of elements in array A after a new entry with value 18 is inserted into this heap? Show all your work. $A = 19, 17, 14, 8, 7, 9, 3, 2, 1, 4$ (8 points)

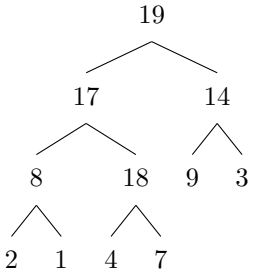
Solution.
We will convert the array representation of the heap to a tree representation of the heap. From here, we can easily trace what happens when 18 is inserted. Then, we can convert the resulting tree back into its corresponding array form to arrive at a final answer.
The tree representation of the initial array is:



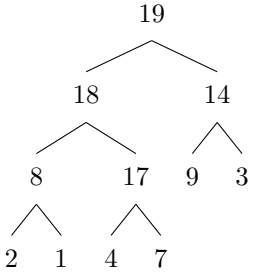
Since the heap's tree representation must always be complete, immediately after inserting 18, the tree representation looks as follows:



In order to maintain the Max-Heap property, the 18 must now be promoted upwards until it is smaller than its parent.
After the first such promotion, the tree looks as follows:



18 is still greater than its parent, 17, so we promote again to get:



Now, 18 is smaller than its parent, 19, so the tree once again satisfies the Max-Heap property. Thus, the process of inserting 18 is finished, so we can convert the resulting tree back to an array. The array that corresponds to the tree after the process of inserting 18 finishes is $A = 19, 18, 14, 8, 17, 9, 3, 2, 1, 4, 7$.
Thus, the order of elements in array A after a new entry with value 18 is inserted into this heap is $A = 19, 18, 14, 8, 17, 9, 3, 2, 1, 4, 7$.

3.

(a) Consider the problem of making change for n cents using the fewest number of coins. Describe a greedy algorithm to make change consisting of quarters (25 cents), dimes (10 cents), nickels (5 cents), and pennies

(1 cent). Prove that your algorithm yields an optimal solution. (Hints: consider how many pennies, nickels, dimes, and dimes plus nickels are taken by an optimal solution at most.)
(b) For the previous problem, give a set of coin denominations for which the greedy algorithm does not yield an optimal solution. Assume that each coin's value is an integer. Your set should include a penny so that there is a solution for every value of n. (15+5 points)

Solution. (a) Our algorithm will implement the Greedy strategy of always using the largest coin possible. We will first pick the largest coin which is $\leq n$. Then, we will decrement n by the value of that coin, and repeat the process until $n = 0$. We will keep track of the frequency with which each coin is used, and return an object that stores the frequency for each of the four coin denominations. The algorithm works as follows:

```
bestChange(int n)
    result.quarterCount = 0
    result.dimeCount = 0
    result.nickelCount = 0
    result.pennyCount = 0
    while (n! = 0)
        if(n ≥ 25)
            result.quarterCount++
            n− = 25
        endIf
        else if(n ≥ 10)
            result.dimeCount++
            n− = 10
        endElseIf
        else if(n ≥ 5)
            result.nickelCount++
            n− = 5
        endElseIf
        else if(n ≥ 1)
            result.pennyCount++
            n− = 1
        endElseIf
    endWhile
    return result
endBestChange
```

Proof of Correctness:
First, we must show that our algorithm always returns a valid combination of coin frequencies, for all positive integers n. Our algorithm runs for the duration of a while loop that terminates when $n = 0$. During each iteration, n decrements by at least 1. When $n \geq 25$, each iteration decrements n by exactly 25 until $n \leq 24$. At this point, each iteration decrements n by exactly 10 until $n \leq 9$. At this point, each iteration decrements n by exactly 5 until $n \leq 4$. Then, each iteration decrements n by exactly 1 until $n = 0$, which is guaranteed to happen in ≤ 4 iterations. Thus, the algorithm will terminate for all $n \in \mathbb{N}$.
Each time the algorithm decrements n by k it increments the count of the coin with value k by 1. Since this process stops when $n = 0$, and we always decrement n to exactly 0, the combination of coin frequencies returned by our algorithm is always valid change for n cents, for all $n \in \mathbb{N}$.

Now, we must show that our algorithm always returns valid change for n cents using the fewest total coins.
Assume to the contrary that our algorithm returns change using more than the fewest possible coins. This means there must be a way to substitute a susbet of coins from our solution for a smaller subset of coins

with the same total value. There are exactly 4 situations which allow for a substitution that decreases the total number of coins used without changing the total value of the coins:
i) $pennyCount \geq 5$ (We can replace 5 pennies with 1 nickel, saving 4 coins)
ii) $nickelCount \geq 2$ (We can replace 2 nickels with 1 dime, saving 1 coin)
iii) $dimeCount \geq 3$ (We can replace 3 dimes with 1 quarter and 1 nickel, saving 1 coin)
iv) $dimeCount \geq 2$ && $nickelCount \geq 1$ (We can replace 2 dimes and 1 nickel with 1 quarter, saving 2 coins)

Thus, since our solution is not optimal, it must allow for one of these situations.
i) pennnyCount ≥ 5 implies that pennyCount was incremented ≥ 5 times. However, pennyCount is only incremented when $0 \leq n \leq 4$, and n decrements by 1 each time pennyCount increments, so pennyCount cannot possibly increment ≥ 5 times. This is a contradiction.
ii) nickelCount ≥ 2 implies that nickelCount was incremented ≥ 2 times. However, nickelCount is only incremented when $5 \leq n \leq 9$, and n decrements by 5 each time nickelCount increments, so nickelCount cannot possibly increment ≥ 2 times. This is a contradiction.
iii) dimeCount ≥ 3 implies dimeCount was incremented ≥ 3 times. However, dimeCount only increments when $10 \leq n \leq 24$, and n decrements by 10 every time dimeCount increments, so dimeCount cannot possibly increments ≥ 3 times. This is a contradiction.
iv) dimeCount ≥ 2 && nickelCount ≥ 1 implies that nickelCount incremented at least once after dimeCount incremented at least twice. However, dimeCount only increments when $10 \leq n \leq 24$, and n decrements by 10 each time dimeCount increments, so $n \leq 4$ after dimeCount increments at least twice. However, nickelCount only increments when $5 \leq n \leq 9$, so nickelCount can never increment after dimeCount increments at least twice. This is a contradiction.

Thus, there is no way to substitute a set of x coins from the solution returned by **bestChange()** for a set of $y < x$ coins that have the same total value.
Thus, the solution returned by **bestChange()** uses the fewest total coins of any combination of coin frequencies that makes valid change for n cents, for all $n \in \mathbb{N}$.
Thus, **bestChange()** always yields the optimal solution.

(b) Consider the set of coin denominations $S := \{1, 3, 10, 11\}$. With an input of $n = 13$, the greedy strategy employed by **bestChange()** would first choose 1 11-cent coin, then 2 1-cent coins, for a total of 3 coins. However, we could also choose 1 10-cent coin and 1 3-cent coin to form 13 cents with just 2 coins. Thus, with this input and this set of coin denominations, the solution returned by the greedy strategy does not have the fewest possible number of coins. Thus, these conditions represent an example in which the greedy strategy does not yield an optimal solution.

4.

You are given positions of N Mice and positions of n holes on a 1-dimensional number line. Each hole can accommodate at most 1 mouse. A mouse can stay in place, move one step right from x to $x+1$, or move one step left from x to $x-1$. Devise an algorithm to assign mice to holes so that the number of moves taken by the mice is minimized. Your algorithm should return the minimum number of moves taken to assign mice to holes and be in $O(NlogN)$ time. (10 points).

Solution.
Note: Based on an instructor-endorsed Piazza post, I assume that n (# of holes) $= N$ (number of mice) for the duration of my solution.
Our algorithm will implement the Greedy strategy of pairing the mouse at the $k'th$ largest position with the hole at the $k'th$ largest position. This should minimize the sum of the absolute differences between hole and mouse positions. Since the absolute difference between the mouse position and hole position is the number of steps the mouse needs to take to get to that hole, this should also minimize the total number of moves taken by the mice. Our algorithm works as follows:

```
assignMice(mice[n], holes[n])
    Sort mouse positions in mice[n] in ascending order
    Sort hole positions in holes[n] in ascending order
    count = 0
    for i : 1 → n
        assign mice[i] to holes[i]
        count + = |mice[i] − holes[i]|
    endFor
    return count
endAssignMice
```

Proof of Correctness:
First, we must show our algorithm never assigns more than one mouse to one hole. Since each mouse $mice[i]$ is assigned to a distinct hole $holes[i]$, we know that each hole is assigned exactly 1 mouse, so no hole is assigned more than 1 mouse.
Now, we must show that our algorithm returns an accurate number of moves needed to assign each $mice[i]$ to the corresponding $holes[i]$. For each i, after $mice[i]$ is assigned to $holes[i]$, count increments by the absolute value of $mice[i] - holes[i]$. Since each move taken by $mice[i]$ moves it 1 step closer to $holes[i]$, this is accurately updating the number of moves taken for each assignment. Thus, our algorithm returns the correct number of moves needed to assign each $mice[i]$ to the corresponding $holes[i]$.
Now, we must prove that our algorithm always returns the minimum number of moves needed. Since our algorithm accurately counts the number of moves needed to produce the assignment it simulates, we just need to show that our assignment of mice to holes is optimal.
Assume to the contrary that there exists an optimal solution O which takes fewer moves than the one produced by **assignMice()**. This means there is at least one *inversion* in the optimal solution at which $mice[i] > mice[j]$, $holes[i] > holes[j]$, but $mice[i]$ is assigned to $holes[j]$, and $mice[j]$ is assigned to $holes[i]$. There are 6 potential orderings $< a, b, c, d >$ of $mice[i], mice[j], holes[i], holes[j]$ under these conditions. We will show that, with each of these orderings, reversing the inversion can only reduce the total number of moves used by the solution.

1) $mice[i] \geq mice[j] \geq holes[i] \geq holes[j]$:

$$\begin{aligned} \implies |mice[i] - holes[j]| + |mice[j] - holes[i]| = & \quad mice[i] - holes[j] + mice[j] - holes[i] \\ = & \quad (mice[i] - holes[i]) + (mice[j] - holes[j]) \\ = & \quad |mice[i] - holes[i]| + |mice[j] - holes[j]| \end{aligned}$$

So removing this type of inversion does not change the total number of moves used.

2) $mice[i] \geq holes[i] \geq mice[j] \geq holes[j]$:

$$\begin{aligned} \implies |mice[i] - holes[j]| + |mice[j] - holes[i]| \geq & \quad |mice[i] - holes[j]| \\ \geq & \quad |mice[i] - holes[j]| + mice[j] - holes[i] \\ = & \quad mice[i] - holes[j] + mice[j] - holes[i] \\ = & \quad |mice[i] - holes[i]| + |mice[j] - holes[j]| \end{aligned}$$

So removing this type of inversion can only decrease or not change the total number of moves used.

3) $mice[i] \geq holes[i] \geq holes[j] \geq mice[j]$:

$$\begin{aligned} \implies |mice[i] - holes[j]| + |mice[j] - holes[i]| = & \quad mice[i] - holes[j] + holes[i] - mice[j] \\ \geq & \quad mice[i] - holes[i] + holes[j] - mice[j] \\ = & \quad |mice[i] - holes[i]| + |mice[j] - holes[j]| \end{aligned}$$

So removing this type of inversion can only decrease or not change the total number of moves used.

4) $holes[i] \geq holes[j] \geq mice[i] \geq mice[j]$:

$$\begin{aligned} \implies |mice[i] - holes[j]| + |mice[j] - holes[i]| &= holes[j] - mice[i] + holes[i] - mice[j] \\ &= holes[i] - mice[i] + holes[j] - mice[j] \\ &= |mice[i] - holes[i]| + |mice[j] - holes[j]| \end{aligned}$$

So removing this type of inversion does not change the total number of moves used.

5) $holes[i] \geq mice[i] \geq holes[j] \geq mice[j]$:

$$\begin{aligned} \implies |mice[i] - holes[j]| + |mice[j] - holes[i]| &= mice[i] - holes[j] + holes[i] - mice[j] \\ &\geq holes[i] - mice[i] + holes[j] - mice[j] \\ &= |mice[i] - holes[i]| + |mice[j] - holes[j]| \end{aligned}$$

So removing this type of inversion can only decrease or not change the total number of moves used.

6) $holes[i] \geq mice[i] \geq mice[j] \geq holes[j]$:

$$\begin{aligned} \implies |mice[i] - holes[j]| + |mice[j] - holes[i]| &= mice[i] - holes[j] + holes[i] - mice[j] \\ &\geq holes[i] - mice[i] + mice[j] - holes[j] \\ &= |mice[i] - holes[i]| + |mice[j] - holes[j]| \end{aligned}$$

So removing this type of inversion can only decrease or not change the total number of moves used.

Therefore, for each inversion present in O, we can remove the inversions 1 by 1 until we have the solution returned by **assignMice()**, and this solution is guaranteed to take $\leq$ as many moves as the optimal solution O. However, we assumed that the O took fewer moves than the solution returned by **assignMice()**. Thus, we have a contradiction, which proves that **assignMice()** always returns the fewest possible number of total moves.

Time Complexity Analysis:
Sorting the list of mice positions take $O(Nlog(N))$ time. Similarly, sorting the list of hole positions takes $O(Nlog(N))$ time. There are N total iterations of the for loop. Inside each iteration, the work done assigning a mouse to a hole (making a pair) and incrementing count takes $O(1)$ time. Thus, the whole for loop takes $O(N)$ time. Therefore, the total runtime of our algorithm is $O(Nlog(N)) + O(Nlog(N)) + O(N) = O(2Nlog(N) + N) = O(Nlog(N))$, as required.

5.

Farmer John has N cows $(1, 2, ..., N)$ who are planning to escape to join the circus. His cows generally lack creativity. The only performance they came up with is the "cow tower". A "cow tower" is a type of stunt in which every cow (except for the bottom one) stands on another cow's back and supports all cows above in a column. The cows are trying to find their position in the tower. Cow I $(i = 1, 2, ..., N)$ has weight W_i and strength S_i. The "risk value" of cow i failing (R_i) is equal to the total weight of all cows on its back minus S_i. We want to design an algorithm to help cows find their positions in the tower such that we minimize the maximum "risk value" of all cows. For each of the following greedy algorithms either prove that the algorithm correctly solves this problem or provide a counter-example.
Hint: One of the two solutions is correct and the other is not.
(a) Sort cows in ascending order of S_i from top to bottom

(b) Sort cows in ascending order of $S_i + W_i$ from top to bottom. (15 points total)

Solution.
(a). This solution does NOT solve this problem.
Counterexample: Let $N = 2$, $C_1 : (S_1 = 100, W_1 = 10)$, $C_2 : (S_2 = 10, W_2 = 500)$. Then sorting cows in ascending order of S_i from top to bottom puts C_1 on the bottom and C_2 on top. C_2 has no cows on top of it, so its risk factor is $R_2 = -S_2 = -10$. C_1 has C_2 on top of it, so its risk factor is $R_1 = W_2 - S_1 = 500 - 100 = 400$. However, if we put C_2 on the bottom and C_1 on top, then C_1's risk factor is $R_1 = -S_1 = -100$ and C_2's risk factor is $R_2 = W_1 - S_2 = 10 - 10 = 0$.
The situation returned by solution (a) has a maximum risk value of 400. The alternative situation has a maximum risk value of only 0. Clearly, solution (a) does not result in a solution with the minimal maximum risk value of all cows.

(b) This solution DOES solve the problem.
Proof. Assume to the contrary that there is a different ordering O that has less maximal risk than the ordering returned by solution (b). Then some C_i must be above some C_j such that $S_i + W_i \geq S_j + W_j$. As we go up from C_j to C_i, there must be at least one *inversion* at which $S_k + W_k \leq S_{k+1} + W_{k+1}$ (with subscripts now denoting a cow's position from the bottom of the tower).
At this inversion, we know $R_k = W + W_{k+1} - S_k$ and $R_{k+1} = W - S_{k+1}$, where W is the weight of all cows above C_{k+1}.
If we flip the inversion, we only change the risk factors of the two adjacent cows, so let's examine how those risk factors change. We have $R_k^* = W + W_k - S_{k+1}$ and $R_{k+1}^* = W - S_k$.
Clearly, before flipping the inversion, since $R_k - R_{k+1} = W_{k+1} + S_{k+1} - S_k \geq W_{k+1} + S_{k+1} - (S_k + W_k) \geq 0$, so the maximum risk of the two relevant cows is R_k.
After flipping the inversion, $R_k - R_k^* = W_{k+1} + S_{k+1} - (W_k + S_k) \geq 0$, and $R_k - R_{k+1}^* = W_{k+1} \geq 0$. Thus, flipping the inversion can only decrease or not change the maximum risk value of all cows. Therefore, we can flip inversions 1 by 1 until we obtain the cow tower returned by solution (b), and this tower is guaranteed to have $\leq$ the maximum risk value from the optimal solution. However, we assumed that the optimal solution O has less maximal risk than the ordering returned by solution (b), so we have a contradiction.
Thus, we have proven that solution (b) always minimizes the maximum risk value of all cows in the cow tower.

Assignment 4

1.

[10 points] Design a data structure that has the following properties (assume n elements int he data structure, and that the data structure properties need to be preserved at the end of each operation:

- Find median takes $O(1)$ time
- Inser takes $O(logn)$ time

Do the following:
(a) Describe how your data structure will work.
(b) Give algorithms that implement the Find-Median() and Insert() functions.

Solution.
(a) We will use two heaps to implement our data structure. One heap will be a max heap that stores the smallest half of the elements and the other heap will be a min heap that stores the largest half of the elements.
Let the elements in our data structure be denoted by the set $S := \{s_1, ..., s_n\}$ such that $s_i \leq s_j$ for all $i \leq j$.
If n is even, our data structure will maintain the property that both heaps have the same number of elements

after each operation. In this way, the maximum element in the max heap will be $s_{\frac{n}{2}}$ and the minimum element in the min heap will be $s_{\frac{n}{2}+1}$. In any ordered list of n elements $s_1, ..., s_n$ where n is even, the median of the elements is

$$\frac{s_{\frac{n}{2}} + s_{\frac{n}{2}+1}}{2}$$

Since we can get the min element from the min heap ($s_{\frac{n}{2}+1}$) and the max element from the max heap ($s_{\frac{n}{2}}$) in O(1) time, maintaining our data structure like this should allow for calculating the median in $O(1)$ time if n is even.
If n is odd, our data structure will maintain the property that the min heap has one more element than the max heap. This ensures that $s_{\lceil \frac{n}{2} \rceil}$ is the minimum element in the min heap. In any ordered list of n elements $s_1, ..., s_n$ where n is odd, the median of the elements is just $s_{\lceil \frac{n}{2} \rceil}$. Since we can get the min element from the min heap ($s_{\lceil \frac{n}{2} \rceil}$) in $O(1)$ time, maintaining our data structure like this should allow for calculating the median in $O(1)$ time if n is odd.
Thus, for all n, our data structure should calculate the median in $O(1)$ time.

For inserting an element into our data structure, we will always insert the first element into the min heap, to ensure that we maintain the aforementioned property when n is odd. For each subsequent insert, we will check if that element is greater than the minimum element in the min heap. If so, we will just insert it into the min heap. Otherwise, we will insert it into the max heap. Afterwards, we check if the number of elements in the max heap is greater than the number of elements in the min heap. If so, we remove the max element from the max heap and insert it into the min heap, which ensures we maintain the aforementioned property for when n is odd. Otherwise, we check if the difference between the number of elements in the min heap and the number of elements in the max heap is greater than 1. If so, we remove the minimum element from the min heap and insert it into the max heap, which ensures we maintain the aforementioned property when n is even.
Thus, in each insert, we maximally do two heap inserts and 1 heap removal, all of which are $O(logn)$ operations, for a total of $O(logn)$ total runtime. All comparison operations are O(1), so the insertions and removals dominate the runtime of the function. Thus, the total runtime of our insert function should be $O(logn)$ for all n.

(b) Note: For both algorithms, *largest* refers to the min heap of the data structure's biggest elements, and *smallest* refers to the max heap of the data structure's smallest elements.
We implement Find-Median() as follows:

Find-Median()
    if($smallest.size + largest.size\%2 == 0$)
        return $(smallest.ExtractMax + largest.ExtractMin)/2$
    endIf
    else
        return $largest.ExtractMin$
    endElse
endFind-Median

Time Complexity Analysis:
Since *smallest* and *largest* are both heaps, we know that $smallest.size$ and $largest.size$ are $O(1)$ operations. Also, since *smallest* is a max heap, we know that $smallest.ExtractMax$ is an $O(1)$ operation. Similarly, since *largest* is a min heap, we know that $largest.ExtractMin$ is an $O(1)$ operation. Therefore, Find-Median() takes a constant number of $O(1)$ operations, so it has a total of $O(1)$ time complexity.

We implement Insert() as follows:

Insert(i)
    if($largest.size == 0$)

largest.insert(i)
endIf
else if($i > largest.ExtractMin$)
largest.insert(i)
endElseIf
else
smallest.insert(i)
endElse
if($smallest.size > largest.size$)
temp = smallest.ExtractMax
smallest.RemoveMax
largest.insert(temp)
endIf
else if($largest.size - smallest.size > 1$)
temp = largest.ExtractMin
largest.RemoveMin
smallest.insert(temp)
endElseIf
endInsert

Time Complexity Analysis:
In the worst case, there are 2 heap inserts and 1 heap removal for a single insertion into our data structure. These each take $O(logn)$ time, so the total runtime is $O(3logn) = O(logn)$. Since the rest of the operations in the algorithm take constant time, we know that Insert() has $O(logn)$ overall time complexity.

Note: Since our algorithm inserts its first element into the *largest* heap, and we always check if $largest.size - smallest.size > 1$ after each insertion, at which point we decrease $largest.size$ by 1 and increase $smallest.size$ by 1, it is impossible for $largest.size - smallest.size > 1$ after an insertion. Similarly, since we always check if $smallest.size > largest.size$ after each insertion, at which point we increment $largest.size$ by 1 and decrement $smallest.size$ by 1, it is impossible for our $smallest.size > largest.size$ after an insertion. Therefore, our insertion method maintains the two necessary properties identified in part (a) for our Find-Median() function to work properly.

2.

[10 points] Let us say that a graph $G = (V, E)$ is a near tree if it is connected and has at most $n + k$ edges, where $n = |V|$ and k is a constant. Give an algorithim with running time $O(n)$ that takes a near tree G with costs on its edges, and returns a minimum spanning tree of G. You may assume that all edge costs are distinct.

Solution. We want to avoid sorting all of the edges in E, as this will take $O((n + k)log(n + k)) = O(nlogn)$ runtime. Our strategy will involve dealing will all of the $n + k$ edges in linear $O(n + k) = O(n)$ time, then dealing with a constant $k + 1$ number of edges in the constant $O((k + 1)log(k + 1)) = O(1)$ time. This should result in $O(n+k+klogk) = O(n)$ total runtime for our algorithm. We will utilize a Union-Find data structure using compression as well as a heap to deal with all of the edges in linear time.
Note: In our algorithm, *biggest* is a heap that stores the $k+1$ edges from E with the heaviest costs, $c(u, v) =$ the cost of a specific edge (u, v), $find(u) =$ the connected component to which a node u belongs.
Our algorithm works as follows:

NearTreeMST(V,E)
$nSmallest$ = cost of the n'th lowest cost edge in E
MakeUnionFind(V)

```
    solution = null
    initialize a min heap called biggest to be empty
    for all edges (u, v) ∈ E
        if(c(u, v) < nSmallest)
            if(find(u)! = find(v))
                add (u, v) to solution
                UnionMerge(find(u), find(v))
            endIf
        endIf
        else
            add (u, v) to biggest
        endElse
    ExtractMin from biggest k + 1 times to get a sorted list of the k + 1 highest
    cost edges in E, called biggestSorted
    for i : 1 → k + 1
        (u, v) = biggestSorted(i)
        if(find(u)! = find(v))
            add (u, v) to solution
            UnionMerge(find(u), find(v))
        endIf
    endFor
    return solution
endNearTreeMST
```

Time Complexity Analysis:
We can find the cost of the n'th lowest cost edge in E in $O(n + k) = O(n)$ time using an algorithm like introselect.
Using a Union-Find data structure with compression like the one described in the textbook, the MakeUnion-Find(V) call only take $O(n)$ runtime.
Also, each of the calls to $find(u)$ or $find(v)$ have an amortized runtime of $O(1)$ using this data structure, and the UnionMerge() calls also have $O(1)$ runtime. There are a constant number of each of these operations during each of the $n + k$ iterations of the first for loop, for a total of $O(n + k)$ runtime.
Each insertion into *biggest* takes no more than $O(log(k + 1))$ time, and there will be $k + 1$ insertions, for a runtime bounded by $O((k + 1)log(k + 1)) = O(klogk)$.
Therefore, over $n + k$ iterations of our first for loop, the UnionFind operations contribute a total of $O(n + k)$ runtime, while the heap insert operations contribute a total of $O(klogk)$ runtime, for a total of $O(n + k + klogk) = O(n)$ runtime, since k is a constant.
Converting *biggest* into *biggestSorted* takes $k + 1$ calls to heap remove, each of which take no more than $O(log(k+1))$ runtime. Therefore, the total runtime of converting *biggest* into *biggestSorted* is upper bounded by $O((k + 1)log(k + 1)) = O(klogk)$.
In the second for loop, each of the operations are just UnionFind operations with $O(1)$ runtime. Therefore, over the $k + 1$ iterations of this for loop, the total runtime is upper bounded by $O(k + 1) = O(k)$.
Adding all of these values together, we can see that the total runtime of the function is bounded by $O(n + n + n + klogk + k) = O(3n + klogk + k) = O(n)$, as required.

Note: Since the MST must have $n - 1$ edges, the cheapest possible MST on a near tree graph would include each of its $n - 1$ cheapest edges. Therefore, if any of these $n - 1$ cheapest edges connects two otherwise disconnected components, it should be in the MST. For the $k + 1$ most expensive edges, these should only be in the MST if they connect two components that are only otherwise connected by more expensive edges. Our algorithm adds all of the $n - 1$ cheapest edges that connect disconnected components to the solution first then adds the $k + 1$ most expensive edges in ascending order of cost, so the solution it returns is a proper minimum spanning tree.

3

[14 points] A new startup FastRoute wants to route information along a path in a communication network, represented as a graph. Each vertex and each edge represent a router and a wire between routes respectively. The wires are weighted by the maximum bandwidth they can suport. FastRoute comes to you and asks you to develop an algorithm to find the path with maximum bandwidth from any source s to any destination t. As you would expect, the bandwidth of a path is the minimum of the bandwidths of the edges on that path; the minimum edge is the bottleneck. Explain how to modify Dijkstra's algorithm to do this.

Solution. We need to make two essential changes to Dijkstra's algorithm to find the path with the maximum bandwidth from any source s to any source t.
First, note that Dijkstra's tries to *minimize* path cost, while the algorithm we want needs to *maximize* the bandwidth of a path. In Dijkstra's we want to find a node $v \notin S$ and an edge (u, v) $(u \in S)$ such that the total cost from s to u added to the cost of edge (u, v) is minimized. In our algorithm, we want to find a node $v \notin S$ and an edge (u, v) $(u \in S)$ such that the minimum of the bandwith of the path from s to u and the bandwith of the edge (u, v) is maximized.

Second, note that in Dijkstra's, the *distance* array holds the sum of the cost of each edge along the cheapest path from the start node s to some node $v \in V$, while our algorithm's *distance* array holds the minimum of the edge costs along the path with the maximal minimum edge cost from s to a node $v \in V$. Thus, instead of updating the distance of a new $v \in S$ by $distance[v] = distance[u] + c(u, v)$, we now must update the distance of a new $v \in S$ by $distance[v] = min(distance[u], c(u, v))$. This ensures our *distance* array properly stores the minimum cost (bandwidth) of the edges in a path from s to a node $v \in V$ instead of the sum of the costs of the edges in that path.

Applying these changes to the Dijkstra's pseudocode found in the textbook, we can write our new algorithm:
Note: For all edges, $b(u, v)$ refers to the bandwidth of that edge. For all nodes, $bandwidth[v]$ refers to the highest bandwidth path from s to v.

maxBandwidth(E,V, s, t)
    Let S = the set of explored nodes
    Let *bandwidth* = the array storing the bandwidth of the path with
    maximum bandwidth from s to each node in S
    Initialize $S = \{s\}$, $bandwidth(s) = \infty$, $s.predecessor = null$
    while $(S! = V)$
        select the node $v \notin S$ and the edge (u, v) $(u \in S)$ such that
        $temp = min(bandwidth[u], b(u, v))$ is maximized.
        Let $v.predecessor = u$
        add v to S
        let $bandwidth(v) = temp$
        if$(v == t)$
            break
        endIf
    endWhile
    Let $temp = t$
    Let $maxBandwidthPath = null$
    while$(temp! = null)$
        add *temp* to $maxBandwidthPath$
        Let $temp = temp.predecessor$

endWhile
return $maxBandwidthPath$
endMaxBandwidth

As you can see, the two main differences between maxBandwidth() and Dijkstra's algorithm are the way the node $v \notin S$ and edge (u, v) $(u \in S)$ are chosen and the way that the *distance* array is updated.

4.

Given a connected graph $G = (V, E)$ with positive edge weights. In V, s and t are two nodes for shortest path computation, prove or disprove with explanation.
(a) If all edge weights are unique, then there is a single shortest path between any two nodes in V.
(b) If each edge's weigth is increased by k, then the shortest path cost between s and t will increase by a multiple of k.
(c) If the weight of some edge e decreases by k, then the shortest path cost between s and t will decrease by at most k.
(d) If each edge's weight is replaced by its square, i.e., w to w^2. then the shortest path between s and t will be the same as before but with different costs.

Solution.
(a) **False.** Consider the graph $G = (V, E)$, where $V = \{s, t, v\}$, and $E = \{(s, v, 1), (v, t, 2), (s, t, 3)\}$. Then all edge weights are unique, as required.
However, the path $p = (s, v, 1), (v, t, 2)$ has the total cost $c(p) = 1 + 2 = 3$. The only other simple path from s to t is $p' = (s, t, 3)$, which also has the total cost $c(p') = 3$. Thus, in this example, there are two distinct shortest paths from s to t, each with cost 3. This counterexample disproves the claim from part (a).

(b) **False.** Consider the graph $G = (V, E)$, where $V = \{s, t, v_1, v_2\}$, and $E = \{(s, v_1, 1), (v_1, v_2, 2), (v_2, t, 3), (s, t, 7)\}$.
Then the shortest path cost from s to t is $p = (s, v_1, 1), (v_1, v_2, 2), (v_2, t, 3)$, which has a total cost of $c(p) = 1 + 2 + 3 = 6$ (this is less than the cost of the only other simple path from s to t $(s, t, 7)$, which has a total cost of 7).
Now, add $k = 5$ to the cost of each edge $e \in E$. Our new graph is $G' = (V, E')$, where $E' = \{(s, v_1, 6), (v_1, v_2, 7), (v_2, t, 8), (s, t, 12)$.
Let's count the total cost of the two simple paths from s to t in our new graph G'. We have $p_1 = (s, v_1, 6), (v_1, v_2, 7), (v_2, t, 8)$, which has a total cost of $c(p_1) = 6 + 7 + 8 = 21$. We also have $p_2 = (s, t, 12)$, which has a total cost of $c(p_2) = 12$. Thus, the shortest cost path in G' has a cost of 12.
However, $12 - 6 = 6$, and 6 is not a multiple of $k = 5$. Therefore, after adding $k = 5$ to each of the edges, the shortest cost path did *not* increase by a multiple of $k = 5$. This counterexample disproves the claim from part (b).

(c) **True**.
Proof. Consider a graph $G = (V, E)$ whose shortest path from s to t is p, where $c(p) = x$.
Now, decrease the weight of some arbitrary $e \in E$ by k to produce a new graph $G' = (V, E')$. Let $p' =$ the shortest path from s to t in G', where $c(p') = y$.
We need to show that $x - y \leq k$.
Case 1: p does *not* include edge e. Since p is the shortest cost path from s to t in G, we know that any path $p*$ from s to t that *includes* edge e must have cost $c(p*) \geq x$. All of these paths $p*$ decrease in cost by exactly k after e is reduced in weight by k. Therefore, after the weight of e is reduced, the cost of all paths $p*$ in G' is $c(p*) - k$. Since $c(p*) \geq x$, we know

$$x - (c(p*) - k) = x - c(p*) + k \leq k$$

If p' *includes* e, then $c(p') = y = c(p*) - k$ for some $p*$. Thus, if p' includes e, we know $x - y \leq k$. If p' does not include e, then the cost of p' is the same in G and G'. Since the cost of all paths in G is $\geq x$ by the

definition of $x = c(p)$, we know that $y = c(p') \geq x$, which implies that $x - y \leq 0 \leq k$.
Thus, if p does *not* include edge e, we know that $x - y \leq k$.

Case 2: p *does* include edge e.
The cost of p after reducing edge e's weight by k is $x - k$. The cost of all other paths from s to t in G are $\geq x$ by the definition of $x = c(p)$. The costs of all paths from s to t that do *not* include e will remain constant after the weight of e is reduced, so the costs of all these paths will still be $\geq x \geq x - k$. The costs of all paths from s to t that *do* include e will decrease by exactly k, so the costs of all these paths will still be $\geq x - k$ since they were $\geq x$ before being reduced by k. Therefore, the costs of all paths from s to t after reducing e's weight by k will still be $\geq x - k$. Thus, the cost of the shortest path from s to t in G' is $c(p') = y \geq x - k$, which directly implies that

$$x - y \leq x - (x - k) = k$$

Thus, if p *does* include edge e, we know that $x - y \leq k$.
Thus, regardless of whether p includes or doesn't include e, we know that $x - y \geq k$ for all $k \in \mathbb{N}$. This completes the proof that reducing one edge weight by k can reduce the shortest path from s to t by at most k.

(d) **False.** Consider the graph $G = (V, E)$, where $V = \{s, t, v_1, v_2\}$, and $E = \{(s, v_1, 1), (v_1, v_2, 2), (v_2, t, 3), (s, t, 5)\}$.
The only two simple paths from s to t are $p_1 = (s, v_1, 1), (v_1, v_2, 2), (v_2, t, 3)$, which has a total cost of $c(p_1) = 1 + 2 + 3 = 6$, and $p_2 = (s, t, 5)$, which has a total cost of $c(p_2) = 5$. Thus, the shortest cost path from s to t just goes through the edge $(s, t, 5)$.
Now, consider the graph $G' = (V, E')$, where $E' = (s, v_1, 1), (v_1, v_2, 4), (v_2, t, 9), (s, t, 25)\}$. Then all edge weights have been squared, as required.
However, the two simple paths from s to t in G' are $p_3 = (s, v_1, 1), (v_1, v_2, 4), (v_2, t, 9)$ with a total cost of $c(p_3) = 1 + 4 + 9 = 14$ and $p_4 = (s, t, 25)$, which has a total cost of $c(p_4) = 25$. Thus, the shortest cost path from s to t in G' goes through edges $(s, v_1, 1)$, $(v_1, v_2, 4)$, and $(v_2, t, 9)$.
Therefore, after squaring each of the edge weights, the shortest path from s to t went through different edges than before squaring the edge weights. This counterexample disproves the claim from part (d).

5.

Consider a directed, weighted graph G where all edge weights are positive. You have one Star, which allows you to change the weight of any one edge to zero. In other words, you may change the weight of any one edge to zero. Propose an efficient method based on $Dijkstra's$ algorithm to find a lowest-cost path from node s to node t, given that you may set one edge weight to zero.

Solution. Since all the edge weights are positive, we know that the lowest-cost path from s to t will inevitably decrease in weight if one of its edges is set to 0. Since setting an edge to 0 is the only way we can modify the graph, we know that the lowest-cost path from s to t, given that one edge weight may be set to zero, must include the edge that is set to 0. Therefore if we can find the lowest-cost path from s to t that goes through edge e when just e has weight 0, for all $e \in E$, the minimum cost path of all of these will be the solution. Therefore, we could solve the problem with brute-force via $m = |E|$ calls to Dijkstra's. For each edge, we would set that edge's weight to 0, run Dijkstra's, store the value of *distance(t)*, then set the edge back to its initial weight. We could then compare the m values of *distance(t)* in linear time to find the maximum of these values, which would have the lowest-cost of any path from s to t, given that we can set one edge weight to 0. However, Dijkstra's runs in $O(mlogn)$, so calling it m times would result in $O(m^2logn)$ runtime, which is not very efficient. We want to calculate the lowest-cost path from s to t through an edge e when just e has weight 0 for all edges e in a constant number of calls to Dijkstra's.
Note: If edge $e = (u, v)$ has weight 0, then the the cost of the lowest-cost path from s to t through e equals the cost of the lowest-cost path from s to u plus the cost of the lowest-cost path from v to t. Therefore, we just need to find the cost of the lowest-cost path from s to v and from v to t for all $v \in V$.

A simple run of Dijkstra's with starting node s will produce an array storing the cost of the lowest-cost path from s to v for all $v \in V$. If we were to flip the direction of all edges $e \in E$, then run Dijkstra's with t as the starting node, we will get an array storing the cost of the lowest-cost path from v to t for all $v \in V$. Thus, we only need to run Dijkstra's twice, keeping track of both cost arrays, and we will have all the information we need to determine which edge in e is set to 0 in the lowest-cost path from s to t. Once we have done this, we can set the weight of e to 0, then run Dijkstra's again on the initial graph with s as the starting node. Based on this third run of Dijkstra's, we can then trace predecessors from t to s to obtain the lowest-cost path from s to t, given that one edge weight can be set to 0. Since we only require 3 calls to Dijkstra's regardles of m, this should be much more efficient than the brute-force method.
We can implement the described method as follows:

```
modifiedDijkstras(s, t, G)
    let G' = a copy of graph G
    run Dijkstra's on G with starting node s
    store path costs in an array called cost1
    Flip the direction of each edge in G'
    run Dijkstra's on G' with t as the starting node
    store path costs in an array called cost2
    let min = ∞
    let minEdge = null
    for each edge (u, v) ∈ E
        if(cost1(u) + cost2(v) < min)
            let min = cost1(u) + cost2(v)
            let minEdge = (u, v)
        endIf
    endFor
    In initial graph, set c(minEdge) = 0
    Run Dijkstra's again on modified initial graph with start node s
    store predecessors of nodes in array called predecessor
    let temp = t
    let path = null
    while(predecessor(temp)! = null)
        add temp to path
        let temp = predecessor(temp)
    endWhile
    return path
endModifiedDijkstras
```

Time Complexity Analysis:
Making a copy of the graph takes $O(m + n)$ time.
Running Dijkstra's the first time takes $O(mlogn)$ time.
Flipping the edge directions in G' takes $O(m)$ time.
Running Dijkstra's the second time takes $O(mlogn)$ time.
Each iteration inside for loop takes constant time, and there are m iterations, so the for loop takes a total of $O(m)$ time.
Running Dijkstra's the third time takes $O(mlogn)$ time.
The while loop can maximally loop through every edge in the graph, and each iteration takes constant time, so the whole while loop takes $O(m)$ time.

Thus, the total runtime of the algorithm is $O((m + n) + mlogn + m + mlogn + m + mlogn + m) = O(n + 4m + 3mlogn) = O(mlogn)$ runtime. This is the same as the asymptotic complexity of Dijkstra's itself, which speaks to the efficiency of this solution.

Assignment 5

1.

[20 points] For the following recurrence equations, solve for $T(n)$ if it can be found using the master method (make sure to show which case applies and why). Else, indicate that the master method is not applicable and explain why.

(a) $T(n) = 8T(n/2) + nlogn - 2023n$

(b) $T(n) = 2T(n/2) + n^3(logn)^3$

(c) $T(n) = 4T(n/2) + n^2(logn)^2$

(d) $T(n) = 3T(n/3) - nlogn$

Solution.

(a) We have

$$T(n) = 8T(n/2) + nlogn - 2023n = aT(n/b) + f(n) \implies a = 8, \quad b = 2, \quad f(n) = nlogn - 2023n$$

Therefore, we can easily see that

$$n^{log_b a} = n^{log_2 8} = n^3$$

Also, since $nlogn > 2023n$ for all $n > 2^{2023}$, we know that

$$f(n) = nlogn - 2023n = \Theta(nlogn)$$

Since $nlogn \leq n^2$ for all $n > 2$, we know that with $\varepsilon = 1$, we have

$$f(n) = \Theta(nlogn) = O(n^{log_b a - \varepsilon}) = O(n^{3-1}) = O(n^2)$$

Therefore, applying *Case 1* of the Master Theorem, we find that

$$T(n) = \Theta(n^{log_b a}) = \Theta(n^3)$$

Thus, the overall asymptotic complexity of $T(n)$ is

$$T(n) = 8T(n/2) + nlogn - 2023n = \Theta(n^3)$$

(b) We have

$$T(n) = 2T(n/2) + n^3(logn)^3 = aT(n/b) + f(n) \implies a = 2, \quad b = 2, \quad f(n) = n^3log^3n$$

We can easily compute that

$$n^{log_b a} = n^{log_2 2} = n^1 = n$$

Also, since $n^3log^3n > n^2$ for all $n > 2$, we know that with $\varepsilon = 1$, we have

$$f(n) = n^3log^3n = \Theta(n^3log^3n) = \Omega(n^2) = \Omega(n^{log_2 2+1}) = \Omega(n^{log_b a+\varepsilon})$$

Applying *Case 3* of the Master Theorem, we find that

$$T(n) = \Theta(f(n)) = \Theta(n^3log^3n)$$

Thus, the overall asymptotic complexity of $T(n)$ is

$$T(n) = 2T(n/2) + n^3(logn)^3 = \Theta(n^3log^3n)$$

(c) We have

$$T(n) = 4T(n/2) + n^2(logn)^2 = aT(n/b) + f(n) \implies a = 4, \quad b = 2, \quad f(n) = n^2log^2n$$

We can easily compute that

$$n^{log_b a} = n^{log_2 4} = n^2$$

Comparing this with the asymptotic complexity of $f(n)$, we find that with $k = 2$, we have

$$f(n) = n^2log^2n = \Theta(n^2log^2n) = \Theta(n^{log_2 4}log^2n) = \Theta(n^{log_b a}log^k n)$$

Applying the special case of *Case 2* of the Master Theorem, we find that

$$T(n) = \Theta(n^{log_b a}log^{k+1}n) = \Theta(n^2log^3n)$$

Therefore, the overall asymptotic complexity of $T(n)$ is

$$T(n) = 4T(n/2) + n^2(logn)^2 = \Theta(n^2log^3n)$$

(d) We have

$$T(n) = 3T(n/3) - nlogn = aT(n/b) + f(n) \implies a = 3, \quad b = 3, \quad f(n) = -nlogn$$

Note: We can interpret $f(n)$ as $f(n) = C(n)+D(n)$, where $D(n)$ is the time needed to divide a problem into subproblems and $C(n)$ is the time needed to combine the results from those subrproblems into a final solution. Therefore, $f(n)$ should always be positive asymptotically, as there is no way to complete the divide and combine steps in negative time. However, $f(n) = -nlogn < 0$ for all $n > 2$, so $f(n)$ does not apply to the situation described by the Master Theorem.
Thus, we cannot apply the Master Theorem for

$$T(n) = 3T(n/3) - nlogn$$

because $f(n)$ is asymptotically negative, so $T(n)$ does not satisfy the conditions of the Master Theorem.

2.

[10 points] Consider the divide and conquer solution described in class to find the closest pair of points in a 2D plane. Assume that we did not have a driver routine to sort the points. So our recursive function did not receive the points in sorted orders of their X and Y coordinates and the sorting had to be done for each subproblem (at every level). What would be the worst-case complexity of this algorithm assuming that the rest of the algorithm remains the same?

Solution.
It is easiest to consider how the recurrence relation changes from the original recurrence relation when we force the recursive function to sort the X and Y coordinates for every subproblem at every level.
For clarity:
Let L = the line dividing the plane.
Let A = the section of the plane containing the leftmost half of the points.
Let B = the section of the plane containing the rightmost half of the points.
In the original solution, for n total points, we divide the the plane into two parts, A and B, which each have approximately $\frac{n}{2}$ points. Since the points are already sorted by x and y coordinates, we can do this in linear $O(n)$ time by selecting the $\lceil \frac{n}{2} \rceil$ points with the lowest x coordinates for one part, leaving the remaining points for the other part. After our recursion returns the closest pair of points in A and the closest pair of points in B, we take the minimum-distance pair. If we let the distance of that pair be x, we just have to do constant work for each of the points within x from the dividing line L to determine if any pairs crossing

L are closer together than x. If we let c be a constant, this leaves us with $cn = O(n)$ work to combine the recursive solutions and find the closest pair of points in the entire plane. This leaves us with a recurrence relation of

$$T(n) = aT(n/b) + D(n) + C(n) = 2T(n/2) + O(n) + O(n)$$

When we force the recursive function to sort the X and Y coordinates for every subproblem at every level, we still need to split the plane into two parts, A and B, each of which have approximately $\frac{n}{2}$ points. However, since the points are not sorted, determining which points to put in A and B cannot be done in linear time. Instead, we must first sort the list of points by X and Y coordinates, which takes a total of $\Theta(nlogn)$ time. After sorting the coordinates, we can find the leftmost half of the points in linear $O(n)$ time just line in the original solution. Therefore, the total runtime of the divide step when forcing the recursive function to sort the points by coordinates is $D(n) = \Theta(nlogn)$. Since we now have lists of the points sorted by X and Y coordinates, when our recursion returns the closest pairs of points in A and B, we can still determine the closest pair of points in the whole plane in $O(n)$ time. Therefore, the recurrence relation when we force the recursive function to sort the points for every subproblem at every level is

$$T(n) = aT(n/b) + D(n) + C(n) = 2T(n/2) + \Theta(nlogn) + O(n)$$

To analyze the worst-case complexity of this recurrence relation, we apply the Master Theorem. We have

$$T(n) = 2T(n/2)+\Theta(nlogn)+O(n) = aT(n/b)+f(n) \implies a = 2, \quad b = 2, \quad f(n) = \Theta(nlogn)+O(n) = \Theta(nlogn)$$

We can easily compute that

$$n^{log_b a} = n^{log_2 2} = n^1 = n$$

Comparing this with the complexity of $f(n)$, we find that, with $k = 1$, we have

$$f(n) = \Theta(nlogn) = \Theta(n^{log_2 2} log^1 n) = \Theta(n^{log_b a} log^k n)$$

Applying the special case of *Case 2* of the Master Theorem, we find that

$$T(n) = \Theta(n^{log_b a} log^{k+1} n) = \Theta(nlog^2 n)$$

Therefore, the worst-case complexity when we force the recursive function to sort the X and Y coordinates for each subproblem at every level is $\Theta(nlog^2 n)$.

3.

[10 points] Solve Kleinberg and Tardos, Chapter 5, Exercise 3.
Suppose you're consulting for a bank that's concerned about fraud detection, and they come to you with the following problem. They have a collection of n bank cards that they've confiscated, suspecting them of being used in fraud. Each bank card is a small plastic object, containing a magnetic stripe with some encrypted data, and it corresponds to a unique account in the bank. Each account can have many bank cards corresponding to it, and we'll say that two bank cards are equivalent if they correspond to the same account.
It's very difficult to read the account number off a bank card directly, but the bank has a high-tech "equivalence tester" that takes two bank cards and, after performing some computations, determines whether they are equivalent.
Their question is the following: among the collection of n cards, is there a set of more than n/2 of them that are all equivalent to one another? Assume that the only feasible operations you can do with the cards are to pick two of them and plug them in to the equivalence tester. Show how to decide the answer to their question with only $O(nlogn)$ invocations of the equivalence tester.

Solution.
Our solution relies on the following observation.

Observation 1: If more than $n/2$ of the cards are equivalent to one another, then when we split the n cards into two equal groups of size $n/2$, at least one of them will have more than $n/4$ cards that are equivalent to one another.

This observation has strong implications for the structure of our recursive Divide and Conquer approach, as determining if there are more than $n/4$ cards equivalent to one another in a group of size $n/2$ is the exact same as determining if more than $n/2$ of the cards in the group of size n are equivalent to one another. This implies we should split our n cards into two groups of size $n/2$ for recursion. For each group, we can use recursion to determine if more than $n/4$ cards are equivalent to one another. If so, we can return one of those cards. If not, we can return null. If both groups return equivalent cards, then we know we have $> n/4 + n/4 = n/2$ cards which are equivalent to one another. If both groups return null, then we know there is no way for more than $n/2$ of the cards to be equivalent to one another by *Observation 1.* If the groups return cards which are not equivalent, then we can check the equivalency of each of the n cards with each of the returned cards to generate two separate counts. If either of these counts is $> n/2$, we can return the card associated with that count. Otherwise, we can return null to indicate that there are *not* more than $n/2$ cards which are equivalent to one another.

Observation 2. If there is a group of only $n = 1$ card, then that card itself is trivially more than $n/2$ cards equivalent to one another.

This implies that the base case of our recursion should trigger when the input list has size $n = 1$, at which point it should return the one bank card in the list.

After we have written the recursive function, to answer the bank's question can simply call our recursive function once on the full list of bank cards. If the return value is not null, we can return true. Otherwise, we return false.

We can implement the algorithm described above as follows, assuming that the "equivalence tester" can be called with equiv(bankcard b1, bankcard b2):

```
majorityEquivalent(list bankcards)
    if( majorityCard(bankcards) != null) return true
    return false
endMajorityEquivalent

majorityCard(list bankcards)
    let n = bankcards.size()
    if (n == 1) return bankcards[1]
    Let first = bankcards[1, ⌈n/2⌉]
    Let last = bankcards[⌈n/2⌉ + 1, n]
    Let majorityFirst = majorityEquivalent(first)
    Let majorityLast = majorityEquivalent(last)
    if (majorityFirst == majorityLast) return majorityFirst
    Let count = 0
    for i: 1 → n
        if (equiv(majorityFirst, bankcards[i]) == true)
            increment count by 1
        endIf
    endFor
    if (count > n/2) return majorityFirst
    Let count = 0
    for i: 1 → n
```

```
        if (equiv(majorityLast, bankcards[i]) == true)
            increment count by 1
        endIf
    endFor
    if (count > n/2) return majorityLast
    return null
endMajorityCard
```

Time Complexity Analysis:
We need to show that our algorithm does $O(nlogn)$ invocations to the "equivalence tester." To do so, we can show that our algorithm as a whole takes $\Theta(nlogn)$ time (assuming $O(1)$ runtime for the "equivalence tester"), as this means that the individual lines calling the "equivalence tester" will never execute more than $cnlogn$ times for some constant c and all $n \geq n_0 > 0$. To analyze the runtime of our algorithm under the above assumption, we can apply the Master Theorem.

We need to express our algorithm as a recurrence relation of the form

$$T(n) = aT(n/b) + D(n) + C(n) = aT(n/b) + f(n) \qquad (1)$$

We split our initial list of n bankcards into two lists of $n/2$ bankcards, so we have $a = 2 = b$. To do so, we only need to calculate the ceiling of $n/2$, which can be done in constant time. Therefore, the asymptotic complexity of the divide step is

$$D(n) = \Theta(1)$$

When we combine the results from our two recursive calls, we use two for loops that each take exactly n iterations. Assuming the calls to the "equivalence tester" take $\Theta(1)$ time, each iteration of each for loop should only take $\Theta(1)$ time, so each foor lop should have a total of $\Theta(n)$ runtime. Therefore, the asymptotic complexity of the combine step is

$$C(n) = \Theta(n) + \Theta(n) = \Theta(2n) = \Theta(n)$$

Plugging these results into (1), we find that

$$T(n) = aT(n/b) + D(n) + C(n) = 2T(n/2) + \Theta(1) + \Theta(n) \implies f(n) = \Theta(1) + \Theta(n) = \Theta(n)$$

We can easily compute that

$$n^{log_b a} = n^{log_2 2} = n^1 = n$$

Comparing this with the asymptotic complexity of $f(n)$, we find

$$f(n) = \Theta(n) = \Theta(n^{log_2 2}) = \Theta(n^{log_b a})$$

Applying *Case 2* of the Master Theorem, we find that

$$T(n) = 2T(n/2) + \Theta(1) + \Theta(n) = \Theta(n^{log_b a} logn) = \Theta(nlogn)$$

Therefore, our algorithm must have $O(nlogn)$ invocations of the "equivalence tester," as required.

4.

[10 points] You are given with two integers a and b, and a variation of Fibonacci series, with $f(0) = a$ and $f(1) = b$. Recall that the Fibonacci sequence is $f(n) = f(n-1) + f(n-2)$. Devise an efficient algorithm to find the n'th Fibonacci number with O(log n) time complexity and prove its time complexity using recurrence relation. (Hint: You can represent the calculation of Fibonacci series using matrix multiplication as follows

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} f(n-1) \\ f(n-2) \end{bmatrix} = \begin{bmatrix} f(n-1) + f(n-2) \\ f(n-1) \end{bmatrix} = \begin{bmatrix} f(n) \\ f(n-1) \end{bmatrix}$$

You can repetitively multiply the resultant matrix with $\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}$ to get subsequent Fibonacci numbers.)

Solution.
Claim: The hint directly implies that

$$\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}^{n-1} \cdot \begin{bmatrix}b\\ a\end{bmatrix} = \begin{bmatrix}f(n)\\ f(n-1)\end{bmatrix}$$

for all $n \geq 2$.

Proof. We apply mathematical induction on n.
Base Case: $n = 2$, we are given that $f(0) = a$ and $f(1) = b$, so plugging in $n = 2$ to the hint equation directly yields

$$\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix} \cdot \begin{bmatrix}b = f(1) = f(n-1)\\ a = f(0) = f(n-2)\end{bmatrix} = \begin{bmatrix}a + b = f(2) = f(n)\\ b = f(1) = f(n-1)\end{bmatrix}$$

so the claim holds for the base case of $n = 2$.
Inductive Hypothesis: Assume that

$$\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}^{n-1} \cdot \begin{bmatrix}b\\ a\end{bmatrix} = \begin{bmatrix}f(n)\\ f(n-1)\end{bmatrix}$$

for all $2 \leq n \leq k$.
Inductive Step: Consider $n = k + 1$. We want to show that

$$\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}^{k} \cdot \begin{bmatrix}b\\ a\end{bmatrix} = \begin{bmatrix}f(k+1)\\ f(k)\end{bmatrix}$$

From the inductive hypothesis, we know that

$$\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}^{k-1} \cdot \begin{bmatrix}b\\ a\end{bmatrix} = \begin{bmatrix}f(k)\\ f(k-1)\end{bmatrix}$$

Premultiplying both sides by $\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}$ yields

$$\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix} \cdot \begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}^{k-1} \cdot \begin{bmatrix}b\\ a\end{bmatrix} = \begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}^{k} \cdot \begin{bmatrix}b\\ a\end{bmatrix} = \begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix} \cdot \begin{bmatrix}f(k)\\ f(k-1)\end{bmatrix} = \begin{bmatrix}f(k) + f(k-1)\\ f(k)\end{bmatrix} = \begin{bmatrix}f(k+1)\\ f(k)\end{bmatrix}$$

which completes the inductive step. The conclusion that the claim holds for all $n \geq 2$ follows by induction. □

We can use this property to calculate $f(n)$ in $O(logn)$ time. Since we can multiply two 2 by 2 matrices together in $O(1)$ time, we only need to compute $\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}^{n-1}$ in $O(logn)$ time, which we can do using recursion. We want to write a recursive function that calculates $\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}^{n-1}$, so we should think about the various cases our function might have to deal with.

If $p = n - 1 \leq 2$, we can directly compute $\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}^{n-1}$ in $O(1)$ through matrix multiplication.

Otherwise, if $p = n - 1$ is even, then $\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}^{n-1} = \left(\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}^{\frac{n-1}{2}}\right)^2$, so we can compute $\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}^{n-1}$ in $O(1)$ time if we first compute $\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}^{\frac{n-1}{2}}$. To do so, we could use recursion with $p = \frac{n-1}{2}$ as the power of $\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}$.

If $p = n - 1$ is odd, then $\begin{bmatrix}1 & 1\\1 & 0\end{bmatrix}^{n-1} = \begin{bmatrix}1 & 1\\1 & 0\end{bmatrix} \cdot \left(\begin{bmatrix}1 & 1\\1 & 0\end{bmatrix}^{\frac{n-2}{2}}\right)^2$, so we can compute $\begin{bmatrix}1 & 1\\1 & 0\end{bmatrix}^{n-1}$ in $O(1)$ time if we first compute $\begin{bmatrix}1 & 1\\1 & 0\end{bmatrix}^{\frac{n-2}{2}}$. To do so, we could use recursion with $p = \frac{n-2}{2}$ as the power of $\begin{bmatrix}1 & 1\\1 & 0\end{bmatrix}$. Therefore, on each recursive call, p decreases to at most $\frac{p}{2}$. Since we only recurse until $p \leq 2$, and dividing p by 2 $logn$ times is guaranteed to result in $p = 1$, we know that we recurse at most $O(logn)$ times. Furthermore, the computation done at each level of recursion is $O(1)$, so the overall complexity of the algorithm should be $O(logn)$.

We could implement the algorithm described above as follows:

realFib(int n, int a, int b)
    if ($n == 0$) **return** a
    if ($n == 1$) **return** b
    Let $arr = \begin{bmatrix}1 & 1\\1 & 0\end{bmatrix}$
    Let $arr = power(n-1, arr)$
    return $b \cdot arr[0][0] + a \cdot arr[0][1]$
endRealFib

power(int p, int[][] arr)
    if ($p == 0 || p == 1$) **return** arr
    if($p\%2 == 0$)
        Let $arr = power(p/2, arr)$
        Let $temp = arr$
        Let $arr[0][0] = temp[0][0] * temp[0][0] + temp[0][1] * temp[1][0]$
        Let $arr[0][1] = temp[0][0] * temp[0][1] + temp[0][1] * temp[1][1]$
        Let $arr[1][0] = temp[1][0] * temp[0][0] + temp[1][1] * temp[1][0]$
        Let $arr[1][1] = temp[1][0] * temp[0][1] + temp[1][1] * temp[1][1]$
    endIf
    else
        Let $arr = power((p-1)/2, arr)$
        Let $temp = arr$
        Let $arr[0][0] = temp[0][0] * temp[0][0] + temp[0][1] * temp[1][0]$
        Let $arr[0][1] = temp[0][0] * temp[0][1] + temp[0][1] * temp[1][1]$
        Let $arr[1][0] = temp[1][0] * temp[0][0] + temp[1][1] * temp[1][0]$
        Let $arr[1][1] = temp[1][0] * temp[0][1] + temp[1][1] * temp[1][1]$
        Let $temp = arr$
        Let $arr[0][0] = temp[0][0] + temp[0][1]$
        Let $arr[0][1] = temp[0][0]$
        Let $arr[1][0] = temp[1][0] + temp[1][1]$
        Let $arr[1][1] = temp[1][0]$
    endElse
    return arr
endPower

Time Complexity Analysis:
We can use the Master Theorem. First, we need to set up a recurrence relation of the form

$$T(n) = aT(n/b) + D(n) + C(n) = aT(n/b) + f(n)$$

to describe the runtime of our Divide & Conquer algorithm. Let $A = \begin{bmatrix}1 & 1\\1 & 0\end{bmatrix}$. Then at each level of recursion,

our power() function calculates A^n by performing one computation of $A^{n/2}$. Therefore, we are dividing our problem into 1 subproblem of size $n/2$, so $a = 1$ and $b = 2$. Determining the size of our one subproblem only requires determining if n is even or odd, which takes $\Theta(1)$ time. Therefore, the complexity of the divide step is $D(n) = \Theta(1)$. Once we obtain the solution to our one subproblem through recursion, we can compute the final result in constant time. Therefore, the complexity of the combine step is also $C(n) = \Theta(1)$. Combining these results, we find that the following recurrence relation describes our algorithm:

$$T(n) = aT(n/b) + D(n) + C(n) = aT(n/b) + f(n) = T(n/2) + \Theta(1) + \Theta(1) \implies f(n) = \Theta(1) + \Theta(1) = \Theta(1)$$

We can easily compute that

$$n^{log_b a} = n^{log_2 1} = n^0 = 1$$

Comparing this with the asymptotic complexity of $f(n)$, we find

$$f(n) = \Theta(1) = \Theta(n^{log_2 1}) = \Theta(n^{log_b a})$$

Applying *Case 2* of the Master Theorem, we find that

$$T(n) = \Theta(n^{log_b a} log n) = \Theta(log n)$$

Therefore, our algorithm has an asymptotic complexity of $\Theta(log n)$, so its runtime is indeed upper bounded by $O(log n)$, as required.

5.

[10 points] You are given a **sorted** array consisting of $k + 1$ values. Only one of the values appears once, and the rest of the k values appear twice. That is, the size of the array is $2k + 1$. Design an efficient Divide and Conquer algorithm for finding which value appears only once. Partial credit (at most 6 points) will be given for non-Divide and Conquer algorithms. Discuss the runtime for your algorithm.

Solution.
Let $n = 2k + 1$. Then we can solve the problem in $O(n)$ time by brute force linear search, comparing each element to its left and right neighbors (since the list is sorted). We want to design a Divide & Conquer algorithm that is more efficient than brute force, so we need to limit our number of subproblems at each level of the recursive tree to 1.
To do so, consider what happens if we examine the middle element in the array, $a(m)$. In $O(1)$ time, we can compare this element to its leftmost neighbor, $a(m-1)$. If the two elements are identical, then we can check if there are an odd number of elements from $a(0)$ to $a(m-2)$. If so, we know it is impossible to create pairs for each of those leftmost $m-1$ elements, so the unique element must be among them. Otherwise, there must be an odd number of elements from $a(m+1)$ to $a(n-1)$, so we know it is impossible to create pairs for each of those rightmost $n-m-1$ elements, so the unique element must be among them. If $a(m) \neq a(m-1)$, we check if $a(m) = a(m+1)$. If so, we check if there are an odd number of elements from $a(m+2)$ to $a(n-1)$. If so, we know it is impossible to create pairs for each of those rightmost $n-m-2$ elements, so the unique element must be among them. Otherwise, there must be an odd number of elements from $a(0)$ to $a(m-1)$, so it is impossible to create pairs for each of the leftmost m elements, so the unique element must be among them. If the $a(m) \neq a(m+1)$ and $a(m) \neq a(m-1)$, then $a(m)$ must be the unique element since the list is sorted. We can continue to recurse on the subset of the array in which we know the unique element exists until the middle element of that subset is the unique element, which is guaranteed to happen for all lists structured as explained in the problem description. We can also stop recursing immediately if the input list has only 1 element, as we know this element must be the unique element.
We could implement the algorithm described above as follows:

```
findHelp(int[] arr)
    return find(arr, 0, arr.size())
```

```
endFindHelp

find(int[] arr, int start, int end)
    Let m = (start + end)/2
    if (end − start == 1) return arr[m]
    if( m > 0&&m < end − 1)
        if (arr[m − 1] == arr[m])
            if ((m − 1)%2 == 1)
                return find(arr, start, m − 1)
            endIf
            else
                return find(arr, m + 1, end)
            endElse
        endIf
        else if (arr[m + 1] == arr[m])
            if((m − 1)%2 == 1)
                return find(arr, m + 2, end)
            endIf
            else
                return find(arr, start, m)
            endElse
        endElseIf
    endIf
    return arr[m]
endFind
```

Time Complexity Analysis:
We will apply the Master Theorem to analyze the asymptotic complexity of our algorithm. We first need to describe our algorithm using a recurrence relation of the form

$$T(n) = aT(n/b) + D(n) + C(n) = aT(n/b) + f(n)$$

During each recursive call on a list of size n, we split our problem into one subproblem of size $\approx \frac{n}{2}$. Therefore, we know that $a = 1$ and $b = 2$. Determining which half of the array to keep searching requires calculating the midpoint of the array and then checking if $a(m)$ is equivalent to $a(m+1)$ or $a(m-1)$. This can all be done in constant time, so the overall complexity of the divide step is $D(n) = \Theta(1)$. We directly return the result of our one subproblem, so the combine step also takes constant time, and its complexity is $C(n) = \Theta(1)$. Combining these results, we can describe our algorithm with the recurrence relation

$$T(n) = aT(n/b) + D(n) + C(n) = aT(n/b) + f(n) = T(n/2) + \Theta(1) + \Theta(1) \implies f(n) = \Theta(1) + \Theta(1) = \Theta(1)$$

We can easily compute that

$$n^{log_b a} = n^{log_2 1} = n^0 = 1$$

Comparing this with the complexity of $f(n)$, we find that

$$f(n) = \Theta(1) = \Theta(n^{log_2 1}) = \Theta(n^{log_b a})$$

Applying *Case 2* of the Master Theorem, we find that

$$T(n) = T(n/2) + \Theta(1) + \Theta(1) = \Theta(n^{log_b a} logn) = \Theta(logn)$$

Thus, the overall complexity of our algorithm is $\Theta(logn)$, so it is indeed asymptotically *more* efficient than the $O(n)$ linear brute force search.

Assignment 6

1.

From the lecture, you know how to use dynamic programming to solve the $0-1$ knapsack problem where each item is unique and only one of each kind is available. Now let us consider the knapsack problem where you have infinitely many items of each kind. Namely, there are n different types of items. All the items of the same type i have equal size w_i and value v_i. You are offered with infinitely many items of each type. Design a dynamic programming algorithm to compute the optimal value you can get from a knapsack with capacity W.

(a) Define (in plain English) subproblems to be solved. (4 pts)

(b) Write a recurrence relation for the subproblems (6 pts)

(c) Make sure you specify

  (a) base case and their values (2 pts)

  (b) where the final answer can be found (1 pt)

(d) What is the complexity of your solution? (2 pts)

Solution.

(a) We want to find the maximum value you can get from a knapsack with capacity W when considering infinitely many items of types $1, \cdots, n$. If $W > w_n$, the maximum value could either include an object of type n, or it could not include an object of type n. If it includes one, it could include another, but the capacity of the knapsack decreases from W to $W - w_n$. If it doesn't include an object of type n, then the capacity stays at W, but we know we have to fill the knapsack will objects of types $1, ..., n-1$. We want to maximize the value the knapsack can hold with capacity W and all n item types, so we need to first compute the maximum values the knapsack can hold with capacity w and k item types for all $0 \leq w \leq W$ and all $0 \leq k \leq n$. These are the subproblems we need to solve.
Let $opt(k, w) :=$ the optimal value you can get from a knapsack with capacity w using objects of types $1, ..., k$. Then we just need to compute $opt(k, w)$ for all $0 \leq k \leq n$, $0 \leq w \leq W$ to solve all the subproblems.

(b) We want to find a recurrence relation for $opt(k, w)$. Consider the collection of items corresponding to $opt(k, w)$. This collection either includes an object of type k, or it doesn't.
If the collection includes an object of type k, then we get v_k value from that object and $opt(k, w - w_k)$ value from the rest of the collection since our capacity decreases by w_k but we can still pick objects from all k item types. Therefore, if the collection includes an object of type k, we know

$$opt(k, w) = v_k + opt(k, w - w_k)$$

If the collection doesn't include an object of type k, then our knapsack's capacity stays at w, but we only consider picking objects of types $1, ..., k-1$. Thus, if the collection doesn't include an object of type k, we know

$$opt(k, w) = opt(k-1, w)$$

Since we don't know whether the collection includes an object of type k or not, we take the maximum of these to values to find

$$opt(k, w) = max(opt(k, max(0, w - w_k)) + v_k, opt(k-1, w)) \qquad (1)$$

This is our recurrence relation for the subproblems.
Note: The $max(0, w - w_k)$ appears because $opt(k, w)$ is only defined for nonnegative k and w, but $w - w_k$ could be negative. If we know the optimal collection includes an object of type k, we know $w - w_k \geq 0$

since the optimal collection cannot exceed the knapsack's capacity. However, since we generally do not know if the collection includes an object of type k, the $max(0, w-w_k)$ appears to ensure we aren't trying to compute invalid values of $opt()$.

(c) (a) The base cases are when $k = 0$ or $w = 0$, or both.
Case 1: $k = 0$, so we have no item types, so we have no items to place in the knapsack, so we cannot put any value in the knapsack, so the optimal value of the knapsack is 0. Thus, $opt(0, w) = 0$ for all w.
Case 2: $w = 0$, so we have no capacity in the knapsack, so we cannot put any items in it without exceeding the capacity, so the optimal value of the knapsack is 0. Thus $opt(k, 0) = 0$ for all k (assuming positive object weights).

(b) The final answer is the maximum value you can get from a knapsack with capacity W using items of types $1, ..., n$, which is stored in $opt(n, W)$. Thus, you can find the final answer by returning/printing $opt(n, W)$.

(d) There are $(n + 1)(W + 1) = \Theta(nW)$ unique subproblems which we must solve. We can create a 2-D $(n + 1)$ by $(W + 1)$ array to store the values of each $opt(k, w)$ as soon as we compute it. By computing the smaller values of $opt(k, w)$ first, we can ensure that we can always compute $opt(k, w)$ using (1) in constant $\Theta(1)$ time. We need a total of $(n + 1) \cdot (W + 1) = \Theta(nW)$ iterations to compute each unique subproblem. Therefore, we have a total of $\Theta(nW)$ constant time iterations in our algorithm, which results in an overall runtime of $\Theta(nW)$.

2.

Solve Kleinberg and Tardos, Chapter 6, Exercise 12.

Suppose we want to replicate a file over a collection of n servers, labeled $S_1, S_2, ..., S_n$. To place a copy of the file at server S_i results in a placement cost of c_i, for an integer $c_i > 0$.

Now, if a user requests the file from server S_i, and no copy of the file is present at S_i, then the servers $S_{i+1}, S_{i+2}, S_{i+3}, ...$ are searched in order until a copy of the file is finally found, say at server S_j, where $j > i$. This results in an access cost of $j - i$. (Note that the lower-indexed servers $S_{i-1}, S_{i-2}, ...$ are not consulted in this search.) The access cost is 0 if S_i holds a copy of the file. We will require that a copy of the file be placed at server S_n, so that all such searches will terminate, at the latest, at S_n.

We'd like to place copies of the files at the servers so as to minimize the sum of placement and access costs. Formally, we say that a configuration is a choice, for each server S_i with $i = 1, 2, ..., n - 1$, of whether to place a copy of the file at S_i or not. (Recall that a copy is always placed at S_n.) The total cost of a configuration is the sum of all placement costs for servers with a copy of the file, plus the sum of all access costs associated with all n servers. Give a polynomial-time algorithm to find a configuration of minimum total cost.

(a) Define (in plain English) subproblems to be solved. (4 pts)

(b) Write a recurrence relation for the subproblems (6 pts)

(c) Make sure you specify

(i) base case and their values (2 pts)

(ii) where the final answer can be found (1 pt)

(d) What is the complexity of your solution? (2 pts)

Solution.

(a) We want to find the the configuration of minimal cost among servers $S_1, S_2, ..., S_n$ where a copy of the file must be placed at S_n.
Consider the optimal configuration, O. The closest server to S_n that also has a copy of the file could be

any server S_i for all $0 \leq i \leq n-1$, where S_0 indicates that S_n is the *only* server with a copy of the file. Let S_k be that server. Then the cost of $S_{k+1}, ..., S_{n-1}$ in O is $\sum_{i=k+1}^{n-1} n - i$, and the cost of $S_1, ..., S_k$ in O is the value of the optimal configuration among servers $S_1, ..., S_k$, where a copy of the file must be placed at S_k. Since k could be any integer from 0 to $n-1$, we must compute the value of the optimal configuration among servers $S_1, ..., S_k$ where a copy of the file must be placed at S_k for all $0 \leq k \leq n-1$. These are the subproblems we need to solve.
Let $opt(i)$ = the cost of the optimal (minimum cost) configuration among servers $S_1, ..., S_i$, where a copy of the file must be placed at S_i.
Then we must compute $opt(i)$ for all $0 \leq i \leq n$ to solve all the subproblems.

(b) We want to find a recurrence relation for $opt(i)$. Consider the optimal configuration corresponding to $opt(i)$. Let S_k be the server closest to S_i in that configuration that has a copy of the file s.t. $k \in \{0, 1, ..., i-1\}$. We have $k = 0$ when there are no servers with lower indexes than i that have copies of the file. Then servers $S_1, ..., S_k$ contribute $opt(k)$ to $opt(i)$, and servers S_j, $k < j < i$ have access costs $i - j$, so they contribute $\sum_{j=k+1}^{i-1} i - j$ to $opt(i)$. Since we must place a copy of the file at server i, we know we S_i contributes c_i to $opt(i)$, regardless of k. Since k can be any element in $\{0, 1, ..., i-1\}$, and $opt(i)$ equals the cost of the minimum cost configuration, we know that

$$opt(i) = c_i + \min_{0 \leq k \leq i-1} (opt(k) + \sum_{j=k+1}^{i-1} i - j)$$

This is a valid recurrence relation for the subproblems, but we can simplify it further.
Note that

$$\sum_{j=k+1}^{i-1} i - j = \quad i - (k+1) + ... + i - (i-1) = 1 + ... + i - k - 1 = \sum_{j=1}^{i-k-1} j = \frac{(i-k-1)(i-k)}{2}$$

$$= \quad \frac{(i-k-1)(i-k)}{2} \frac{(i-k-2)!}{(i-k-2)!} = \frac{(i-k)!}{2!(i-k-2)!} = \binom{i-k}{2}$$

This gives us the simplified equation

$$opt(i) = c_i + \min_{0 \leq k \leq i-1} (opt(k) + \binom{i-k}{2}) \qquad (1)$$

This is the recurrence relation we will use for our subproblems.

(c) (i) The base case is when $i = 0$.
$i = 0$, so we have no servers, so the minimal cost configuration of the servers is 0, so $opt(0) = 0$.

(ii) The final answer is the minimum cost configuration associated with $opt(n)$. We can use a predecessor array to extract the configuration from the $opt(n)$ computation in linear time. When calculating $opt(i)$, each time we find k s.t. $opt(k) + \binom{i-k}{2}$ is minimized, we can set $P(i) = k$, after initializing all values of $P(i)$ to 0. Then, we can create a config() array of size n to store the configuration, with all values initialized to 0 except $config(n) = 1$. Then, we can run through a while loop in linear time, conditioning on $P(n)! \neq 0$. During each iteration, we set $config(P(n)) = 1$, then set $n = P(n)$. This will produce a configuration array in which $config(i) = 1 \iff S_i$ has a copy of the file and $config(i) = 0$ otherwise. The final answer can be found in this configuration array.

(d) We must compute $n = O(n)$ unique subproblems to arrive at a final answer. For each subproblem, we have to loop through $O(n)$ possibilities for k. This gives us $O(n)O(n) = O(n^2)$ iterations to find $opt(n)$. Since we calculate the $opt(i)$ for smaller i first, we can always calculate $opt(k)$ in constant time for each iteration. Also, if we use memoization and store the value of $\binom{i-k}{2}$ in a 2-D array at each iteration, we can use Pascal's Identity to always compute $\binom{i-k}{2}$ in constant time. Therefore, it takes $O(n^2)$ constant time iterations to compute $opt(n)$, for a total of $O(n^2)$ runtime. To create the configuration array only takes one run through a linear for loop with constant time iterations, which is $O(n)$ time. Therefore, the overall runtime of our algorithm is $O(n^2) + O(n) = O(n^2)$.

3.

Given n balloons, indexed from 0 to $n-1$. Each balloon is painted with a number on it represented by array $nums$. You are asked to burst all the balloons. If the you burst balloon i you will get $nums[left] \cdot nums[i] \cdot nums[right]$ coins. Here $left$ and $right$ are adjacent indices of i. After bursting the balloon, the $left$ and $right$ then becomes adjacent. You may assume $nums[-1] = nums[n] = 1$ and they are not real therefore you can not burst them. Design a dynamic programming algorithm to find the maximum coins you can collect by bursting the balloons wisely. Analyze the running time of your algorithm.

(a) Define (in plain English) subproblems to be solved. (4 pts)

(b) Write a recurrence relation for the subproblems (6 pts)

(c) Make sure you specify

(i) base case and their values (2 pts)

(ii) where the final answer can be found (1 pt)

(d) What is the complexity of your solution? (2 pts)

Solution.

(a) We want to find the maximum coins that can be collected by wisely bursting balloons $0 \to n-1$. Define O = the optimal sequence of balloons that maximizes the coins collected. Suppose the last balloon to be popped in O is balloon k where $0 \le k \le n-1$. Balloons $0, ..., k-1$ are never adjacent to any balloons from $k+1, ..., n-1$ when they are burst, and vice versa, so the number of coins O gets for bursting balloons $0, ..., k-1$ must be the maximum number of coins that can be collected by wisely bursting balloons $0, ..., k-1$. Similarly, the number of coins O gets for bursting balloons $k+1, ..., n-1$ must be the maximum number of coins that can be collected by wisely bursting balloons $k+1, ..., n-1$. Since k can be any value from 0 to $n-1$, we need to compute the maximum number of coins that can be collected by wisely busting balloons $i, i+1, ..., j-1, j$ for all $0 \le i, j \le n-1$. These are the subproblems we need to solve.
Let $opt(i, j)$ = the maximum number of coins that can be collected by wisely bursting balloons i to j. Then we need to compute $opt(i, j)$ for all $0 \le i, j \le n-1$ to solve all of the subproblems.

(b) We want to find a recurrence relation for $opt(i, j)$. Consider any subset of balloons $i, ..., j$ where $j \ge i$. Let k = the index of the last balloon popped in the optimal sequence corresponding to $opt(i, j)$. Since k is popped last out of $i, ..., j$, we know it is adjacent to $i-1$ and $j+1$ when it bursts. Therefore, we know k contributes exactly $nums[i-1] \cdot nums[k] \cdot nums[j+1]$ coins to $opt(i, j)$. Balloons $i, ..., k-1$ are never adjacent to any balloons from $k+1, ..., j$ when they are burst, and vice versa, so the number of coins $opt(i, j)$ gets for bursting balloons $i, ..., k-1$ is $opt(i, k-1)$ and the number of coins $opt(i, j)$ gets for bursting balloons $k+1, ..., j$ is $opt(k+1, j)$. Since k can be any value from i to j, and $opt(i, j)$ must be maximal, we know that

$$opt(i, j) = \min_{i \le k \le j} (opt(i, k-1) + opt(k+1, j) + nums[i-1] \cdot nums[k] \cdot nums[j+1])$$

for all $0 \le i \le j \le n-1$. This is the recurrence relation we use to solve our subproblems.

(c) (i) Base Case: $i > j$, so, from balloon i to balloon j, there are no baloons, so there is nothing to burst, so we cannot make any coins. Thus, for all $0 \le j < i \le n-1$, $opt(i, j) = 0$.

(ii) If we use a 2D n by n array to store $opt(i, j)$ for all $0 \le i, j \le n-1$, then $opt[0, n-1]$ will store the maximum coins that can be collected by bursting balloons 0 through $n-1$ wisely, which is all the balloons. Thus, we know that $opt[0, n-1]$ stores our final answer.

(d) We need to calculate $opt(i, j)$ for all $0 \leq i, j \leq n-1$, so there are n^2 distinct subproblems to compute. Since we store values of $opt(i, j)$ in $opt[i][j]$ as we compute them and our recurrence relation for $opt(i, j)$ relies on $opt(i, k-1)$ and $opt(k+1, j)$, which have smaller ranges than $opt(i, j)$, we can iterate in increasing order of $j - i$. Then, for each $opt(i, j)$, we can calculate the necessary $opt(i, k-1)$ and $opt(k+1, j)$ in $\Theta(1)$ time. However, we need to do this for all k s.t. $i \leq k \leq j$, which means we must do $O(n)$ work for each $opt(i, j)$. Since we must do this for n^2 values of $opt(i, j)$, our solution has a total runtime of $O(n^2)O(n) = O(n^3)$, which is polynomial as required.

4.

Suppose you have a rod of length N, and you want to cut up the rod and sell the pieces in a way that maximizes the total amount of money you get. A piece of length i is worth p_i dollars. Devise a Dynamic Programming algorithm to determine the maximum amount of money you can get by cutting the rod strategically and selling the cut pieces.

(a) Define (in plain English) the subproblems to be solved. (4 pts)

(b) Write a recurrence relation for the subproblems. (6 pts)

(c) Using the recurrence formula in part b, write pseudocode to solve the problem. (5 pts)

(d) Make sure you specify

(i) base cases and their values (2 pts)

(ii) where the final answer can be found (1 pt)

(e) What is the complexity of your solution (2 pts)

Solution. Define O_i = a set of cuts that yields the maximum amount of money for a rod of length i.

(a)
For a rod of length N, we could index the rod from 0 (left end) to N (right end). We want to find the total money associated with O_N, so we should consider the elements of O_N. If $O_N \neq \emptyset$, the leftmost (lowest index) element in O_N could be anything from 1 to $N-1$. If we know this leftmost element is $k \in \{1, \cdots, N-1\}$, we can consider the rod as being two rods of length k and $N-k$. Then the total money associated with O_N is the sum of the maximum money you can get by strategically cutting the two rods of length k and $N-k$. Since k can range from 1 to $N-1$, this means we have the following subproblems to solve:
Find the maximum amount of money you can get by cutting a rod of length k strategically and selling the cut pieces for all $0 \leq k \leq N$.
If we let $opt(k) :=$ the maximum amount of money you can get by cutting a rod of length k and selling the cut pieces, then to solve all the subproblems, we just need to find $opt(k)$ for all $0 \leq k \leq N$.

(b)
For any rod of length k, if $O_k = \emptyset$, then $opt(k) = p_k$ since there are no cuts in the optimal solution, so the total money is just the price of a piece of length k. Otherwise, the first (lowest index) cut in the optimal set of cuts, O_k, could occur anywhere from 1 to $k-1$. If it happens at $j \in \{1, \cdots, k-1\}$, then we can split the rod into two rods of size j and $k-j$. At this point, we know that the maximum money from cutting the rod of length k is

$$opt(k) = opt(j) + opt(k-j)$$

This is true for all $1 \leq j \leq k-1$, and we want to maximize $opt(k)$, so if $O_k \neq \emptyset$, we can conclude that

$$opt(k) = \max_{1 \leq j \leq k-1} (opt(j) + opt(k-j))$$

Taking the maximum between this value and the value for when $O_k = \emptyset$, we find

$$opt(k) = \max\big(p_k, \max_{1 \le j \le k-1}(opt(j) + opt(k-j))\big)$$

for all $1 \le k \le N$, which is the recurrence relation for the subproblems.

(c)
We will provide an iterative solution that fills out $opt()$, an array of size $N+1$, in increasing order of index to ensure we can calculate each potential $opt(j) + opt(k-j)$ in constant time. For each $1 \le i \le n$, we will loop through all $1 \le j < i$ to compute $\max\big(p_i, \max_{1 \le j \le i-1}(opt(j) + opt(i-j))\big)$. The algorithm works as follows:

```
fillFindOpt(vector<int> prices, int N)
    Let opt[] = array of size N + 1
    Let opt[0] = 0
    if (N == 0)
        return opt[N]
    endIf
    for i: 1 → n
        opt[i] = prices[i-1]
        for j: 1 → i − 1
            if (opt[j] + opt[i − j] > opt[i])
                opt[i] = opt[j] + opt[i − j]
            endIf
        endFor
    endFor
    return opt[N]
endFillFindOpt
```

(d)

(i) Bases Case: We have two base cases, when $n = 0$ and when $n = 1$.
Case 1: $n = 0$, so our rod has no length, so it is impossible to sell any pieces of our rod, so we cannot make any money, so

$$opt(0) = 0$$

Case 2: $n = 1$. so our rod has length 1, so the only way we can sell pieces of our rod is by selling the one piece of length 1 that makes up the entire rod. Pieces of length 1 have a price of p_1, so we know

$$opt(1) = p_1$$

(ii) The solution is the maximum amount of money we can get by cutting the rod of length N strategically and selling the pieces. Therefore, by definition of $opt(k)$, we know the solution is $opt(N)$, which is stored as the last element in the $opt[]$ array after running **fillFindOpt**. Thus, the final answer can be found at $opt[N]$ after running **fillFindOpt**. This is also the value **fillFindOpt** returns, so we could find the final solution by looking at the return value of **fillFindOpt** with the input N.

(e)
It takes $\Theta(N)$ time to initialize the $opt[]$ array. It takes $\Theta(1)$ time to check if $N = 0$ and return $opt[N]$ if so. Therefore, the algorithm takes $\Theta(N) + \Theta(1) = \Theta(N)$ time before the for loops.
Now, let's examine what happens at each iteration of the innermost for loop. Since we are filling out $opt[]$ in increasing order of indices, we can always calculate $opt[j] + opt[i-j]$ in $\Theta(1)$ time. Therefore, each iteration of the innermost for loop takes $\Theta(1)$ time. There are maximally N iterations of the innermost for loop for each of the N iterations of the outermost for loop, for a total of $O(N) \cdot N = O(N^2)$ constant time iterations. Therefore, the for loops take $O(N^2)$ total time. Thus, the total runtime of our algorithm is $\Theta(N) + O(N^2) = O(N^2)$, so our algorithm runs in polynomial time.

5.

Solve Kleinberg and Tardos, Chapter 6, Exercise 10.

You're trying to run a large computing job in which you need to simulate a physical system for as many discrete steps as you can. The lab you're working in has two large supercomputers (which we'll call A and B) which are capable of processing this job. However, you're not one of the high-priority users of these supercomputers, so at any given point in time, you're only able to use as many spare cycles as these machines have available.

Here's the problem you face. Your job can only run on one of the machines in any given minute. Over each of the next n minutes, you have a "profile" of how much processing power is available on each machine. In minute i, you would be able to run $a_i > 0$ steps of the simulation if your job is on machine A, and $bi > 0$ steps of the simulation if your job is on machine B. You also have the ability to move your job from one machine to the other; but doing this costs you a minute of time in which no processing is done on your job.

So, given a sequence of n minutes, a plan is specified by a choice of A, B, or "move" for each minute, with the property that choices A and B cannot appear in consecutive minutes. For example, if your job is on machine A in minute i, and you want to switch to machine B, then your choice for minute $i+1$ must be move, and then your choice for minute $i+2$ can be B. The value of a plan is the total number of steps that you manage to execute over the n minutes: so it's the sum of a_i over all minutes in which the job is on A, plus the sum of b_i over all minutes in which the job is on B.

The problem. Given values $a_1, a_2, ..., a_n$ and $b_1, b_2, ..., b_n$, find a plan of maximum value. (Such a strategy will be called optimal.) Note that your plan can start with either of the machines A or B in minute 1.

(a) part (a) of the question (4pts)
Show that the following algorithm does not correctly solve this problem, by giving an instance on which it does not return the correct answer

(b) part (b) of the question (answer according to the format below)

(i) Define (in plain English) subproblems to be solved. (4 pts)

(ii) Write a recurrence relation for the subproblems (6 pts)

(iii) Using the recurrence formula in part (ii), write pseudocode to solve the problem. (5 pts)

(iv) Make sure you specify

(i) base cases and their values (2 pts)

(ii) where the final answer can be found (1 pt)

(v) What is the complexity of your solution? (2 pts)

Solution. (a)
We use the following counterexample, with $n = 2$:

	Minute 1	Minute 2
Computer A	10	1
Computer B	1	500

Let's walk through the algorithm presented it part (a) to see show that it fails to find an optimal plan.

In minute 1, we choose A because $a_1 = 10 > 5 = b_1$.
Set $i = 2$
$2 \leq n = 2$
choice in minute $i - 1 = 2 - 1 = 1$ was A
$b_{i+1} = b_3 = undefined \not> a_i + a_{i+1} = a_2 + a_3 = 1 + undefined$ Choose A in minute $i = 2$
Proceed to iteration $i + 1 = 2 + 1 = 3$.
$i = 3 \not\leq n = 2$.
Exit while loop.

Here the presented algorithm chooses computer A for both minute 1 and minute 2. This results in a total of $a_1 + a_2 = 10 + 1 = 11$ steps completed, so this plan, P_A, has a value of 11.
However, if we choose computer B for both minute 1 and minute 2, we would complete $b_1 + b_2 = 1 + 500 = 501$ steps, so this plan, P_B, has a value of 501.
Since $501 > 11$, we know that the presented algorithm did not return the plan that maximized the total value. Therefore, the presented algorithm failed to find an optimal plan.

(b)
We will use the following observations to help us answer this question:
Observation 1: Assuming all given values $a_1, \cdots, a_n$ and $b_1, \cdots, b_n$ are positive, the optimal plan for n minutes will never "move" on the nth minute, as it would always make more sense to stay and gain the positive value b_n or a_n.
Observation 2: The optimal plan through n minutes could end on either computer A or computer B.
Observation 3: Without loss of generality, if the optimal solution through n minutes ends on computer A, then it could be on computer A on the $n-1$'th minute. If it is not, then it must be on "move" during the $n-1$'th minute and be on computer B during the $n-2$'th minute.

(i)
From *Observation 1* and *Observation 2*, we know we need to find the optimal plan that ends on either computer. To do this, we can find the optimal plan that ends on computer A, then compare it to the optimal plan that ends on computer B.
From *Observation 3*, we know that in order to compute the optimal plan ending on each computer at minute n, we first need to compute the optimal plan ending on each computer at minute $n-1$ and minute $n-2$. These are the subproblems we need to solve.
Let $opt(i)$ = the value of the optimal plan ending on either computer at minute i.
Let $optA(i)$ = the value of the optimal plan ending on computer A at minute i.
Let $optB(i)$ = the value of the optimal plan ending on computer B at minute i.
Then we need to compute $optA(i)$ and $optB(i)$ for all $1 \leq i \leq n$ to solve all of the subproblems. From here, we can simply set

$$opt(i) = max(optA(i), optB(i)) \qquad (1)$$

to find the value of the optimal plan ending on either computer at minute i for all $1 \leq i \leq n$. This will complete solving the subproblems.

(ii)
For all $2 \leq i \leq n$, we know we can can compute $opt[i]$ by first computing $optA[i]$ and $optB[i]$.
Case 1: $optA[i]$. We know the plan corresponding to $optA[i]$, which we'll call O_A ends on computer A at minute i. By *Observation 3*, either O_A is on computer A at minute $i-1$, or O_A is on "move" at minute $i-1$ and on computer B at minute $i-2$. Among these two possible plans, O_A corresponds to the one with maximum value. Therefore, we know

$$optA[i] = max(optA[i-1] + a_i, optB[i-2] + a_i) \qquad (2)$$

Case 2: $optB[i]$. Similarly, we know the plan corresponding to $optB[i]$, which we'll call O_B ends on computer B at minute i. By *Observation 3*, either O_B is on computer B at minute $i-1$, or O_B is on "move" at minute $i-1$ and on computer A at minute $i-2$. Among these two possible plans, O_B corresponds to the one with maximum value. Therefore, we know

$$optB[i] = max(optB[i-1] + b_i, optA[i-2] + b_i) \qquad (3)$$

We will use (2) and (3) as our two recurrence relations for the subproblems.

(iii)

Using (1), (2) and (3), our algorithm becomes a simple for loop with constant time iterations.
We initialize $optA[0] = optB[0] = 0$ because the value of any plan that lasts 0 minutes has to be 0.
We initialize $optA[1] = a_1, optB[1] = b_1$ because there is only one option for a one minute plan that ends on computer A and there is only one option for a one minute plan that ends on computer B.
We initialize $opt[1] = max(optA[1], optB[1])$ following (1).
Then we just run through all $2 \leq i \leq n$, setting $optA[i]$, $optB[i]$, and $opt[i]$ following (1), (2) and (3). Our algorithm works as follows:

findOptimalPlan($\{a_1, ..., a_n\}$, $\{b_1, ..., b_n\}$)
    let $optA$ = new array of size $n + 1$
    let $optB$ = new array of size $n + 1$
    let opt = new array of size $n + 1$
    let $optA[0] = optB[0] = opt[0] = 0$
    let $optA[1] = a_1$, $optB[1] = b_1$, $opt[1] = max(optA[1], optB[1])$
    for i: $2 \rightarrow n$
        $optA[i] = max(optA[i-1] + a_i, optB[i-2] + a_i)$
        $optB[i] = max(optB[i-1] + b_i, optA[i-2] + b_i)$
        $opt[i] = max(optA[i], optB[i])$
    endFor
    return $opt[n]$
endFindOptimalPlan

(iv)

(i) Base Cases: $n = 1$ and $n = 0$.
Case 1: $n = 0$, so the plan must take 0 minutes, so there is no way for it to complete any number of steps, so the plan must have a value of 0. Thus, $optA[0] = optB[0] = opt[0] = 0$.
Case 2: $n = 1$, so the plan must take 1 minute, so if it ends on A, its value must just be a_i. Otherwise, its value must just be b_i. The optimal plan out of these two options is the one corresponding to the maximum of these two values. Therefore, $optA[1] = a_1$, $optB[1] = b_1$, and $opt[1] = max(optA[1], optB[1])$.

(ii) The final answer is the value of the optimal solution that ends at minute n, which is stored in $opt[n]$. This is also returned by **findOptimalPlan**, so the final answer can be found by looking at the return value of **findOptimalPlan**.

(v)
It takes $\Theta(n)$ time to initialize each of the three size $n + 1$ arrays, for a total of $\Theta(n) + \Theta(n) + \Theta(n) = \Theta(3n) = \Theta(n)$ runtime before the for loop.
Inside the for loop, since we compute the needed $optA[i-1], optA[i-2], optB[i-1], optB[i-2]$ before entering iteration i, we can complete iteration i in $\Theta(1)$ time for all $2 \leq i \leq n$. There are $n - 1$ total constant time iterations, for a total runtime of $\Theta(n)$ inside the for loop.
This leaves **findOptimalPlan** with a total runtime of $\Theta(n) + \Theta(n) = \Theta(2n) = \Theta(n)$, so our algorithm runs in linear time.

Assignment 7

1.

Solve Kleinberg and Tardos, Chapter 6, Exercise 5:

As some of you know well, and others of you may be interested to learn, a number of languages (including Chinese and Japanese) are written without spaces between the words. Consequently, software that works with text written in these languages must address the *word segmentation problem*—inferring likely boundaries

between consecutive words in the text. If English were written without spaces, the analogous problem would consist of taking a string like "meetateight" and deciding that the best segmentation is "meet at eight" (and not "me et at eight," or "meet ate ight," or any of a huge number of even less plausible alternatives). How could we automate this process?

A simple approach that is at least reasonably effective is to find a segmentation that simply maximizes the cumulative "quality" of its individual constituent words. Thus, suppose you are given a black box that, for any string of letters $x = x_1x_2...x_k$, will return a number $quality(x)$. This number can be either positive or negative; larger numbers correspond to more plausible English words. (So $quality($"me"$)$ would be positive, while $quality($"ght"$)$ would be negative.)

Given a long string of letters $y = y_1y_2...y_n$, a segmentation of y is a partition of its letters into contiguous blocks of letters; each block corresponds to a word in the segmentation. The total quality of a segmentation is determined by adding up the qualities of each of its blocks. (So we'd get the right answer above provided that $quality($"$meet$"$) + quality($"at"$) + quality($"$eight$"$)$ was greater than the total quality of any other segmentation of the string.)

Give an efficient algorithm that takes a string y and computes a segmentation of maximum total quality. (You can treat a single call to the black box computing $quality(x)$ as a single computational step.)

(a) Define (in plain English) the subproblems to be solved. (4 pts)

(b) Write a recurrence relation for the subproblems. (6 pts)

(c) Using the recurrence formula in part b, write pseudocode to find the maximum total quality among all segmentation possibilities. (5 pts)

(d) Make sure you specify

(i) base cases and their values (2 pts)

(ii) where the final answer can be found (1 pt)

(e) What is the complexity of your solution? (2 pts)

Solution.
(a)
Suppose y consists of n letters. Then we want to find the segmentation of maximum total quality among these n letters. We can define a segmentation based on the indices of the last letters in each segment. For example, a segmentation $\{n\}$ would just be the entire string y, and a segmentation $\{1, 2, ..., n\}$ would be the string y cut into n 1-letter long segments. In this way, $n \in$ every segmentation. If the segmentation of maximal quality is *not* $\{n\}$, then there exists some $k < n$ s.t. k is the index of last letter of the second to last segment. The segmentation of maximal quality then becomes
{ segmentation of maximal quality for $y_1...y_k\} \cup \{n\}$
Thus, we can find the segmentation of maximal quality for $y_1...y_n$ by first computing the segmentation of maximal quality for $y_1...y_k$ for all $1 \leq k < n$.
Let $OPT(k)$ be the quality of the segmentation of maximal quality for $y_1...y_k$. Then the subproblems we need to solve are $OPT(k)$ for all $1 \leq k \leq n$.
(b)
We want to find a recurrence relation for the quality of the segmentation of maximal quality for $y_1...y_k$.
This could be the entirety of $y_1...y_k$, which has $quality(y_1...y_k)$ quality.
If it isn't, then there exists some $j < k$ such that j is the index of the second to last segment in the segmentation of maximal quality for $y_1...y_k$. In this case, the quality of the segmentation of maximal quality for $y_1...y_k$ is the sum of the maximal quality for $y_1...y_j$ and $quality(y_{j+1}...y_k)$. Since j could be any number from $1 \to k-1$, or the segmentation of maximal quality could exist of just $y_1...y_k$ itself, and we know this segmentation has maximal quality, we know that

$$OPT(k) = \max\Big(quality(y_1...y_k), \max_{j=1\to k-1}(OPT(j) + quality(y_{j+1}...y_k))\Big)$$

This is true for all $k > 1$, so this is the recurrence relation we can use to solve our subproblems.
(c)
Our algorithm will use two for loops, one to loop through the values of k, and one to loop through the values of j for each value of k. We use 1-based indexing in our pseudo-code, and we use y_i to refer to the i'th letter in the input string, as in parts (a) and (b).

FindMaxQualitySegment(y)
    if ($y ==$"") return 0
    let OPT be an array of size $y.size()$
    let $OPT[1] = quality(y_1)$
    For (k: $2 \rightarrow y.size()$)
        let $OPT[k] = quality(y_1...y_k)$
        For (j: $1 \rightarrow k-1$)
            let $temp = OPT[j] + quality(y_{j+1}...y_k)$
            if($temp > OPT[k]$)
                let $OPT[k] = temp$
            endIf
        endFor
    endFor
    return $OPT[y.size()]$
EndFindMaxQualitySegment

(d)

i. The base cases are when $k = 0$ or $k = 1$. When $k = 0$, we have the empty string, which has no quality, so we return 0, and we have $OPT(0) = 0$. When we have $k = 1$, the only segment is the one letter string itself, so $OPT(1) = quality(y_1)$, and we return $quality(y_1)$.

ii. The final answer is the quality of the maximal quality segmentation for $y_1...y_n$, which is stored in $OPT(n)$. Thus, the final answer can be found in $OPT(n)$, which is also the return value of our FindMaxQualitySegment function.

(e)
It takes $\Theta(1)$ time to check if y is the empty string, return 0 if so, and set $OPT[1] = y_1$ if not. It takes $\Theta(n)$ time to create the OPT array. There are $O(n)$ iterations of the outer for loop. For each of these $O(n)$ iterations, there are $O(n)$ iterations of the inner for loop. Inside each inner iteration, we only add, compare, and update values in $\Theta(1)$ time (since we compute $OPT(k)$ for smaller k first), so each iteration takes constant time. Therefore, there are $O(n) \cdot O(n)$ constant time iterations, for a total of $O(n^2)$ runtime for the for loops. Therefore, FindMaxQualitySegment takes a total of $\Theta(1)+\Theta(n)+O(n^2) = O(n^2)$ runtime.

2.

[20 points] You are given an integer array $a[1], ..., a[n]$, find the contiguous subarray (containing at least one number) which has the largest sum and only returns its sum. The optimal subarray is not required to return or compute. Taking $a = [5, 4, -1, 7, 8]$ as an example: the subarray [5] is considered as a valid subarray with sum 5, though it only has one single element; the subarray $[5, 4, -1, 7, 8]$ achieves the largest sum 23; on the other hand, $[5, 4, 7, 8]$ is not a valid subarray as the numbers 4 and 7 are not contiguous.

(a) Define (in plain English) the subproblems to be solved. (4 pts)

(b) Write a recurrence relation for the subproblems. (6 pts)

(c) Using the recurrence formula in part b, write pseudocode to find the subarray (containing at least one number) which has the largest sum. (5 pts)

(d) Make sure you specify

(i) base cases and their values (2 pts)

(ii) where the final answer can be found (1 pt)

(e) What is the complexity of your solution? (2 pts)

Solution.
(a)
We want to find the subarray of a with the largest sum, which we'll call the optimal subarray. Since the optimal subarray must have at least one number, it must end on an element of a. Suppose the optimal subarray ends on the k'th element of a, where $1 \leq k \leq n$. Since the optimal subarray is continuous, it either equals $[a[k]]$, or it includes $a[k-1]$. If it includes $a[k-1]$, then since the optimal subarray has the largest sum, it must include the subarray with the largest sum that ends on $a[k-1]$. To determine whether the optimal subarray equals $[a[k]]$ or includes $a[k-1]$, we must compare the sums of the two corresponding arrays. Since k could be any value from 1 to n, we need to calculate the sum of the optimal subarray ending on k for all $1 \leq k \leq n$. Once we have done so, we can run through the array of these results in linear time, and we will find the subarray with the largest sum.
Let $OPT(k)$ be the sum of the optimal subarray ending on $a[k]$. Then we need to find $OPT(k)$ for all $1 \leq k \leq n$ to solve the subproblems.
(b)
We want to find a recurrence relation for $OPT(k)$. If the optimal subarray ending on $a[k]$ is $[a[k]]$, then $OPT(k) = a[k]$. Otherwise, we know the optimal subarray includes $a[k-1]$. For the optimal subarray ending on $a[k]$ to be optimal, there cannot be a subarray ending on $a[k]$ with a larger sum, so since it includes $a[k-1]$, it must also include all other $a[j]$ that correspond to the array whose sum is $OPT(k-1)$. Since the optimal subarray ending on $a[k]$ must also include $a[k]$, we know its total sum is

$$OPT(k) = a[k] + OPT(k-1)$$

To determine whether or not the optimal array ending on $a[k]$ includes $a[k-1]$, we compare these two values of $OPT(k)$ to find

$$OPT(k) = \max(a[k], a[k] + OPT(k-1))$$

This is the recurrence relation we will use to solve our subproblems.
(c)
We will use the recurrence relation from part (b) to calculate and store $OPT(k)$ for all $1 \leq k \leq n$. However, as we calculate each $OPT(k)$, we will also calculate the length of the corresponding optimal subarray. After doing so, we can run through the $OPT()$ array in linear time, find the value of the optimal subarray, and then use its length to find the subarray itself. We will use 1-based indexing in our pseudocode.

FindMaxSumSubarray(a)
```
let n = a.size()
Let OPT be an array of size n
Let lengths be an array of size n
Let OPT[1] = a[1]
Let lengths[1] = 1
For k: 2 → n
    if (a[k] > a[k] + OPT(k − 1))
        Let OPT[k] = a[k] + OPT(k − 1)
        Let lengths[k] = lengths[k − 1] + 1
    endIF
    else
        Let OPT[k] = a[k]
        Let lengths[k] = 1
    endElse
```

```
    endFor
    let maxIndex = −1
    let max = INTMIN
    For (i: 1 → n)
        if( opt[i] > max)
            Let max = opt[i]
            Let maxIndex = i
        endIf
    endFor
    Let solution be an array of size lengths[maxIndex]
    For (i: maxIndex − lengths[maxIndex] + 1 → maxIndex)
        Add a[i] to solution
    endFor
    return solution
endFindMaxSumSubarray
```

(d)

i. The base case is when $a.size() = n = 1$, at which point $OPT(n) = OPT(1) = a[1]$, and the subarray with the largest sum (that contains at least one element) is $[a[1]]$.

ii. The final answer is the subarray with the largest sum among all nonempty subarrays of a. The sum of the subarray with the largest sum among all nonempty subarrays of a is stored in $opt[maxIndex]$. The optimal subarray that corresponds to $opt[maxIndex]$ is stored in *solution*, and is returned by our function. Thus, the final answer can be found in *solution* and in the return value of $FindMaxSumSubarray$.

(e)
It takes $\Theta(n)+\Theta(n) = \Theta(2n) = \Theta(n)$ time to create the *lengths* and OPT arrays of size n. It takes constant $\Theta(1)$ time to initialize n, $OPT[1]$, and $lenghts[1]$. Therefore, it takes a total of $\Theta(1) + \Theta(n) = \Theta(n)$ time to do the initial steps in FindMaxSumSubarray before the for loops. The first for loop has $\Theta(n)$ iterations, each of which take constant time since we can access $a[k]$ in constant time and we calculate $OPT[k]$ for smaller k first. Therefore, the first for loop has $\Theta(n)$ total runtime. The second for loop also has $\Theta(n)$ constant time iterations for a total of $\Theta(n)$ runtime. The third for loop has $O(n)$ constant time iterations (since we just add an element to the back of an array each iteration) for a total of $O(n)$ runtime. Adding up the runtimes of each consecutive component of FindMaxSumSubarray, we find that the algorithm has $\Theta(n) + \Theta(n) + \Theta(n) + O(n) = \Theta(n)$ total runtime.

3.

[20 points] You are given an array of positive numbers $a[1], ..., a[n]$. For a subarray sequence $a[i_1], a[i_2], .., a[i_t]$ of array a (that is $i_1 < i_2 < ... < i_t$): if it is an increasing sequnce of numbers, that is, $a[i_1], a[i_2], ..., a[i_t]$, its happiness score is given by

$$\sum_{k=1}^{t} k \cdot a[i_k]$$

Otherwise, the happiness score of this array is zero.
For example, for the input $a = [22, 44, 33, 66, 55]$, the increasing subsequence $[22, 44, 55]$ has happiness score $1\cdot22+2\cdot44+3\cdot55 = 275$; The increasing subsequence $[22, 33, 55]$ has happiness score $1\cdot22+2\cdot33+3\cdot55 = 253$; the sequence $[33, 66, 55]$ has happiness score 0 as this sequence is not increasing. Please design an efficient algorithm to **only** return the highest happiness score over all the subsequences.

(a) Define (in plain English) the subproblems to be solved. (4 pts)

(b) Write a recurrence relation for the subproblems. (6 pts)

(c) Using the recurrence formula in part b, write pseudocode to find the highest happiness score over all the subsequences. (5 pts)

(d) Make sure you specify

(i) base cases and their values (2 pts)

(ii) where the final answer can be found (1 pt)

(e) What is the complexity of your solution? (2 pts)

Solution.
(a)
Suppose the sequence corresponding to the maximum happiness score ends on $a[k]$, where $1 \leq k \leq n$. Then this sequence must have some length l where $1 \leq l \leq k$. Therefore $a[k]$ contributes exactly $a[k] \cdot l$ to the happiness score of this optimal sequence ending at $a[k]$. The rest of the happiness score must come from $a[i]$'s where $1 \leq i < k$ and $a[i] < a[k]$, since the optimal sequence must be increasing. To find which $a[i]$ precedes $a[k]$ in the optimal sequence, we must calculate the maximum happiness score of a sequence of length $l-1$ ending at $a[i]$, for all i s.t. $1 \leq i < k$ and $a[i] < a[k]$. Since k can be any value from $1 \to n$, we must calculate the maximum happiness score of a sequence ending at $a[k]$ of length l for all $1 \leq k \leq n$ and all $1 \leq l \leq k$.
Let $OPT(k, l)$ be the maximum happiness score of a sequence ending at $a[k]$ with length l. Then we just need to find $OPT(k, l)$ for all $1 \leq k \leq n$ and all $1 \leq l \leq k$ to solve our subproblems.
(b)
We want to find a recurrence relation for $OPT(k, l)$. Consider the sequence corresponding to $OPT(k, l)$. If $l > 1$, then we know there are multiple elements in the sequence, so we can let $a[j]$ be the element in the sequence that precedes $a[k]$, where $1 \leq j < k$. Then we know $a[j] < a[k]$, and for the sequence to correspond to $OPT(k, l)$, the maximum happiness score of a length $l-1$ sequence ending at $a[j]$ must be greater than the maximum happiness score of a length $l-1$ sequence ending at $a[i]$ for all $i \neq j$ s.t. $i < k$ and $a[i] < a[k]$. Since $a[k]$ itself contributes $l \cdot a[k]$ to $OPT(k, l)$, this yields

$$OPT(k, l) = \max_{j<k s.t. a[j]<a[k]} (OPT(j, l-1) + l \cdot a[k])$$

This is true for all $2 \leq l \leq k \leq n$, and it is the recurrence relation we will use to solve our subproblems.
(c)
We create an n by n array to store the values of $OPT(k, l)$. We then iterate through all values of $OPT(k, l)$, starting with $l = 1$ since our recursive call decreases the value of l. Each time we set a new value of $OPT(k, l)$, we check to see if we need to update the maximum happiness score, which we return at the end of our function. We use 1 based indexing for our pseudocode.

MaxHappiness(a)
```
let n = a.size()
let OPT be an n by n array
let max = −1
let maxLength = −1
let maxIndex = −1
for (l: 1 → n)
    for (k: 1 → n)
        if( l > k)
            OPT[k][l] = 0
        endIf
        else if(l == 1)
            OPT[k][l] = a[k]
        endElseIf
        else
```

```
            let tempmax = −1
            for (i: 1 → k)
                if(a[i] < a[k] && l − 1 ≤ i)
                    if(tempmax < (l · a[k] + OPT[i][l − 1])
                        let tempmax = l · a[k] + OPT[i][l − 1]
                    endIf
                endIf
            endFor
        endElse
        if(opt[k][l] > max)
            let max = opt[k][l]
            let maxIndex = k
            let maxLength = l
        endIf
    endFor
  endFor
  return max
endMaxHappiness
```

(d)

i. Case 1: $l = 1$, so the length of $OPT(k, l)$ is 1, and the sequence ends on $a[k]$, so its only element must be $a[k]$, so its happiness score must be $OPT(k, 1) = a[k]$ for all $1 \leq k \leq n$.
Case 2: $l > k$, so it is impossible to have a sequence of length l that ends at element $a[k]$, as the first element would have to be $a[k - l + 1]$, but $k - 1 + 1 <= 0$ if $l > k$, and $a[1]$ is the first element in a. Thus, $OPT(k, l) = 0$ for all $l > k$.

ii. The final answer is the maximum happiness score of any sequence of length l ($1 \leq l \leq n$) ending at $a[k]$ ($1 \leq k \leq n$). Therefore, the final answer is the maximum value of all $OPT(k, l)$. Since we update the maximum value of all $OPT(k, l)$ each time we calculate a new $OPT(k, l)$, we know this value is stored in the max variable from $MaxHappiness$, which is also the return value of the function. Therefore, the final answer can be found in the max variable or in the return value of $MaxHappiness$. This is also stored in $OPT[maxIndex][maxLength]$.

(e)
It takes $\Theta(1)$ time to declare and initialize n, $maxLength$, max, and $maxIndex$. It takes $\Theta(n^2)$ time to initialize the OPT array of size n by n. This yields a total of $\Theta(1) + \Theta(n^2) = \Theta(n^2)$ runtime before the for loops.
The first for loop has $\Theta(n)$ iterations. The second for loop also has $\Theta(n)$ iterations. For each iteration inside the second for loop, the iteration either takes constant time or $\Theta(k)$ time, both of which are $O(n)$. Therefore, we have a total of $\Theta(n)\Theta(n) = \Theta(n^2)$ $O(n)$ iterations in our for loops, for a total of $\Theta(n^2)O(n) = O(n^3)$ runtime.
Thus, the total runtime of $MaxHappiness$ is $O(n^3)$.

4.

[20 points] You've started a hobby of retail investing into stocks using a mobile app, RogerGood. You magically gained the power to see N days into the future and you can see the prices of one particular stock. Given an array of $prices$ of this particular stock, where $prices[i]$ is the price of a given stock on the i'th day, find the maximum profit you can achieve through various buy/sell actions. RogerGood also has a fixed fee per transaction. You may complete as many transactions as you like, but you need to pay the transaction fee for each transaction (only pay once per pair of buy and sell). Assume you can own at most one unit of stock.

(a) Define (in plain English) the subproblems to be solved. (4 pts)

(b) Write a recurrence relation for the subproblems. (6 pts)

(c) Using the recurrence formula in part b, write pseudocode to solve the problem. (5 pts)

(d) Make sure you specify

(i) base cases and their values (2 pts)

(ii) where the final answer can be found (1 pt)

(e) What is the complexity of your solution? (2 pts)

Solution.
We suppose the fixed transaction fee is t for this problem.
(a)
Since there are N total days, and each transaction (buy then sell) takes two days, there are a maximum of $\lfloor \frac{N}{2} \rfloor$ transactions in the sequence of transactions that maximizes profits. Let's call this sequence the optimal sequence. Then the optimal sequence is the same as the sequence that maximizes profits with at most $\frac{N}{2}$ transactions through the first N days. This sequence either includes a transaction on the N'th day, or it doesn't. If it doesn't, then the maximum profits over the optimal sequence through N days with at most $\frac{N}{2}$ transactions equals the maximum profits over the optimal sequence through $N-1$ days with at most $\frac{N}{2}$ interactions. If the optimal sequence does include a transaction on the N'th day, then it must be a sell, since buying on the last day would not increase profits, so there is some $1 \leq m < N$ s.t. buying on day m maximizes profits. The profits contributed by the first $m-1$ days must then be the maximum profits over the first $m-1$ days with at most $\frac{N}{2}-1$ transactions. Therefore, in order to calculate the maximum profits over all N days with at most $\frac{N}{2}$ transactions, we must first calculate the maximum profits over the first k days with at most j transactions for all $1 \leq k \leq N$, $1 \leq j \leq \frac{N}{2}$.
Let $OPT(k, j)$ be the optimal solution over the first k days with at most j transactions. Then we need to compute $OPT(k, j)$ for all $1 \leq k \leq N$, $1 \leq j \leq \frac{N}{2}$ to solve the subproblems.
(b)
We want to find a recurrence relation for $OPT(k, j)$. Note that a transaction is either made on day k, or it isn't. If it is not, then the maximum profits are $OPT(k-1, j)$, since the profits do not change on day k. If a transaction is made on day k, then it must be a sell, so there is some $1 \leq i < k$ s.t. there is a buy on day i which corresponds to the sell on day k. This contributes profits of $prices[k] - prices[i] - t$ for days i through k. In order to maximize the profits from buying on day i and correspondingly selling on day k, days 1 through $i-1$ must produce the maximum profits from the first $i-1$ days with at most $j-1$ transactions. Since i could be any value from 1 to $k-1$, if there is a sell on day k, then

$$OPT(k, j) = \max_{1 \leq m < k} (prices[k] - prices[i] - t + OPT(i-1, j-1))$$

Since we do not know whether selling on day k maximizes profits, we compare this value with the value when we don't sell on day k to fin

$$OPT(k, j) = \max(OPT(k-1, j), \max_{1 \leq i < k} \Big(prices[k] - prices[i] - t + OPT(i-1, j-1)) \Big)$$

This is the recurrence relation we will use to solve the subproblems.
(c)
We will loop through each value of k in increasing order, looping through each value of $1 \leq j \leq \frac{n}{2}$ for each value of k. Inside each iteration, we will compute the i for which $prices[k] - prices[i] - t + OPT(i-1, j-1)$ is maximal. Then, we will compare this value to $OPT(k-1, j)$ and set $OPT(k, j)$ equal to the maximum of the two. We use 1 based indexing in the pseudocode.

MaxProfits(prices, t)

```
    let n = prices.size()
    let OPT be an n by n/2 array
    if(n == 0) return 0
    for(j: 1 → n/2)
        let OPT[1][j] = 0
    endFor
    for(k: 2 → n)
        for(j: 1 → n/2)
            let OPT[k][j] = OPT[k − 1][j]
            for(i: 1 → k − 1)
                let temp = prices[k] − prices[i] − t + OPT[i − 1][j − 1]
                if( temp > OPT[k][j])
                    let OPT[k][j] = temp
                endIf
            endFor
        endFor
    endFor
    return OPT[n][n/2]
endMaxProfits
```

(d)

i. Case 1: $N = 0$, so there are no days to make transactions, so there is no way to make any profit, so the maximum profit is 0, so we return 0.
Case 2: $N = 1$, so we cannot make any transactions that make any profit (we could just sell and buy at the same price), so the maximum profit we can make is 0, so we let $OPT[1][j] = 0$ for all j.

ii. The final answer is the maximum profit that can be obtained with any number of transactions through all N days. This is equivalent to the maximum profit that can be obtained with at most $\frac{N}{2}$ transactions through all N days. This value is stored in $OPT[N][\frac{N}{2}]$, which is what our function returns. Therefore, the final answer can be found both in $OPT[N][\frac{N}{2}]$ and in the return value of $MaxProfit$.

(e)
It takes $\Theta(N)\Theta(\frac{N}{2}) = \Theta(\frac{N^2}{2}) = \Theta(N^2)$ time to create the OPT array. It takes $\Theta(\frac{N}{2}) = \Theta(N)$ time to initialize $OPT[1][j]$ to 0 for all j. There are two $\Theta(N)$ nested for loops, with a total of $\Theta(N)\Theta(N) = \Theta(N^2)$ iterations. Inside each iteration, we do $O(N)$ constant time iterations through. Therefore, the nested for loops are upper-bounded by $\Theta(N^2)O(N) = O(N^3)$ runtime. This gives a total runtime of $\Theta(N^2)+\Theta(N)+O(N^3) = O(N^3)$ for our MaxProfits algorithm.

Assignment 8

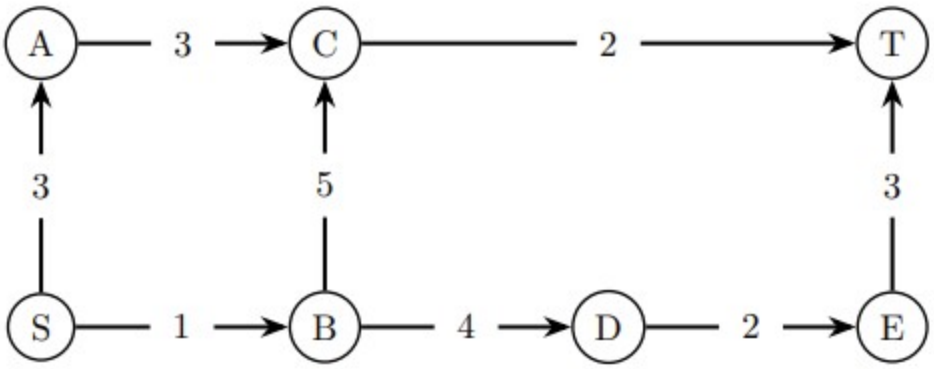
1.

(10pts) The following graph G has labeled nodes and edges between it. Each edge is labeled with its capacity.

(a) Draw the final residual graph G_f using the Ford-Fulkerson algorithm corresponding to the max flow. Please do NOT show all intermediate steps.

(b) What is the max-flow value?

(c) What is the min-cut?

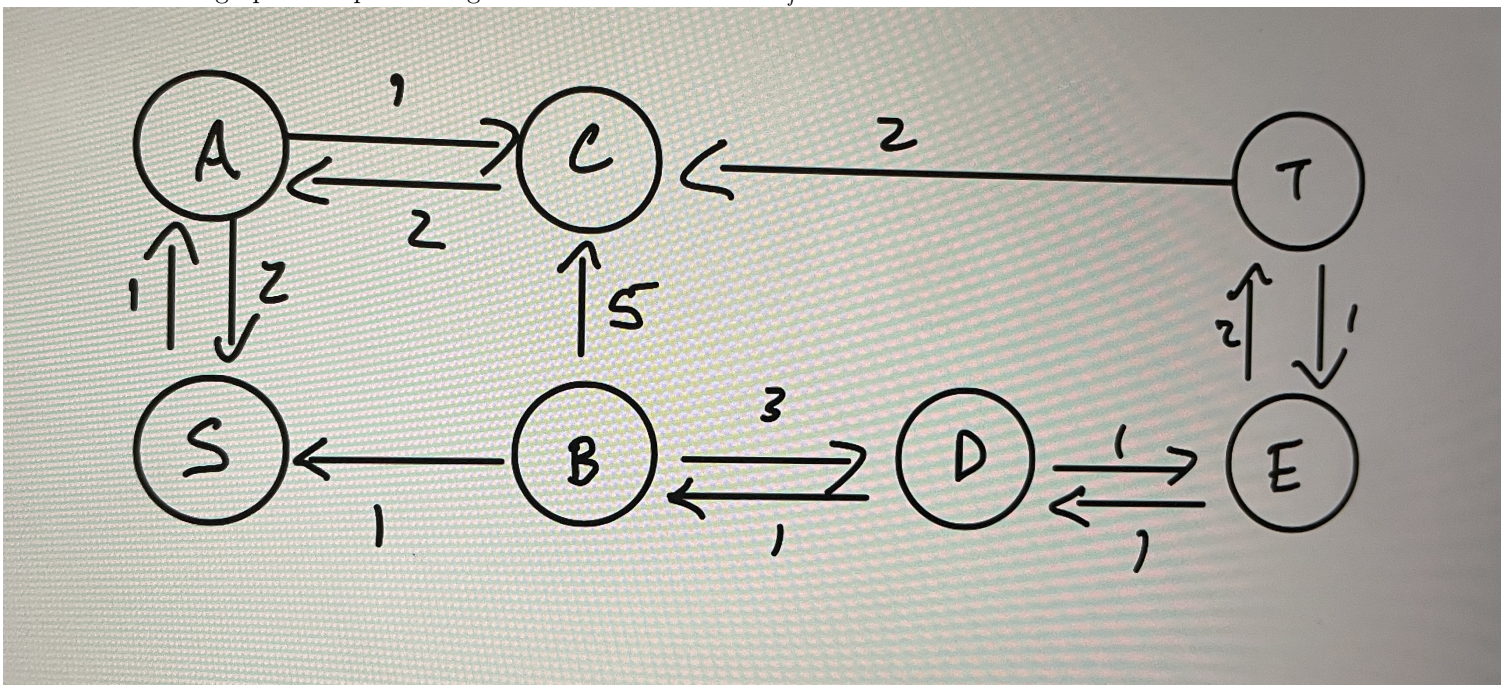


Solution.

(a)

Note: We are using the representation of residual graphs that uses reverse edges, as discussed in lecture.

The final residual graph after performing Ford Fulkerson on G is G_f :



(b)

The max-flow value is equal to the sum of the edges going into the source node, S, in the final residual graph, G_f. Therefore, the value of the max-flow of G is

$$v(f) = 2 + 1 = 3$$

(c)

The min cut is the partitioning of the nodes in $V = \{S, A, B, C, D, E, T\}$ into two disjoint sets, X and Y, where $S \in X$ and $T \in Y$ for which the sum of the edge weights going from X to Y equals $v(f) = 3$.

If we let $X = \{S, A, C\}$ and $Y = \{B, D, E, T\}$, then the only two edges going from X to Y in G are $(C, T, 2)$ and $(S, B, 1)$. The sum of the weights of these edges is $2 + 1 = 3 = v(f)$. Therefore, the min-cut of G is

$$(X, Y) = (\{S, A, C\}, \{B, D, E, T\})$$

2.

(15pts) Determine if the following statements are true or false. For each statement, briefly explain your reasoning.

(a) In a flow network, the value of flow from S to T can be higher than the maximum number of edge disjoint paths from S to T. (Edge disjoint paths are paths that do not share any edge)

(b) For a flow network, there always exists a maximum flow that doesn't include a cycle containing positive flow.

(c) If you have non-integer edge capacities, then you cannot have an integer max flow.

(d) Suppose the maximum s-t flow of a graph has value f. Now we increase the capacity of every edge by 1. Then the maximum s-t flow in this modified graph will have a value of at most $f + 1$.

(e) If all edges are multiplied by a positive number k, then the min-cut remains unchanged.

Solution.
(a)
True. Consider the simple graph $G = (V, E)$, where $V = \{S, A, T\}$ and $E = \{(S, A, 2), (A, T, 2)\}$. Then the only edge disjoint path from S to T is $S- > A- > T$, so the maximum number of edge disjoint paths from S to T is 1. However, the maximum flow from S to T is 2, as the bottleneck of the $S- > A- > T$ path is 2. Since $2 > 1$, the value of the flow from S to T is greater than the maximum number of edge disjoint paths from S to T. This example proves the existence of a flow from S to T that exceeds the maximum number of edge disjoint paths from S to T.

(b)
True. Assume to the contrary that there exists a flow network for which all maximum flows include a cycle containing positive flow. Consider one such maximum flow, M. Consider any s-t path that goes through a cycle with positive flow at least once in M. If the bottleneck of that path is found in the cycle, we can create a path with a higher bottleneck by avoiding the cycle. This will increase the flow of that path without changing the flow of any other paths, thus increasing the overall flow of the graph. However, this contradicts the assumption that M is a maximum flow. If the bottleneck of the path is not found in the cycle, we can create a path with the same bottleneck by avoiding the cycle. This will not change the flow of any path in the graph, so it will not change the overall flow of the graph. We can do this for every s-t path that goes through a positive cycle, and the overall flow of the resulting graph, M', will equal the flow from M. However, M' now has no paths which go through positive cycles. Therefore, M' doesn't include a cycle containing positive flow, but it has the same flow as M. This contradicts the assumption that all maximum flows include a cycle containing positive flow. By contradiction, this concludes the proof that there always exists a maximum flow that doesn't include a cycle containing positive flow for any flow network. For any maximum flow that includes a cycle with positive flow, we can just remove the cycle(s) with positive flow to create a maximum flow that doesn't include a cycle with positive flow.

(c)
False. Consider the simple graph $G = (V, E)$, where $V = \{S, A, B, T\}$ and $E = \{(S, A, 0.5), (S, B, 0.5), (A, T, 0.5), (B, T, 0.5)\}$. Then all edges have non-integer capacities. However, using the two paths $S- > A- > T$ and $S- > B- > T$, which both have a bottleneck of 0.5, we find that the maximum flow of this graph is $v(f) = 0.5+0.5 = 1 \in \mathbb{Z}$. Therefore, although this graph has all non-integer edge capacities, the max flow value, 1, is still an integer. This counterexample disproves the claim that if you have non-integer edge capacities, you cannot have an integer max flow.

(d)
False. Consider the same graph G as in part (c). As discussed in part (c), the max flow of G is $v(f) = f = 1$. Now, let's increase the capacity of each edge by one. This results in the graph $G' = (V, E')$ where

$V = \{S, A, B, T\}$ and $E = \{(S, A, 1.5), (S, B, 1.5), (A, T, 1.5), (B, T, 1.5)\}$. Once again, let's use the paths $S- > A- > T$ and $S- > B- > T$ to construct our max flow for G'. Both of these paths now have a bottleneck of 1.5, so the maximum flow of G' is $v(f) = 1.5 + 1.5 = 3 = f + 2 > f + 1 = 1 + 1 = 2$. Therefore, after increasing all edge capacities in G by 1, the max flow in the modified graph is greater than $f + 1$. This counterexample disproves the claim that increasing the capacity of every edge in a graph by 1 increases the max flow of the graph by at most 1.

(e)
True. Assume to the contrary that there exists a graph for which multiplying all edges by k changes the min-cut. Suppose the initial graph is G with max flow f and min-cut (X, Y), where $c_G(X, Y) = f$. Suppose after multiplying all edges by k the graph is G' with min-cut (X', Y'), where $(X', Y') \neq (X, Y)$ and (X, Y) is not a min-cut of G'. Since (X, Y) is the min-cut in G, we know the capacity of (X', Y') in G is $\geq f = c_G(X, Y)$. Suppose (X, Y) has edges with capacities $w_1, ..., w_n$ going from X to Y in G and (X', Y') has edges with capacities $v_1, ..., v_m$ going from X' to Y' in G. Then we know

$$f = c_G(X, Y) = w_1 + \cdots + w_n \leq v_1 + \cdots v_m = c_G(X', Y')$$

After multiplying all edge capacities by k, we have

$$c_{G'}(X, Y) = kw_1 + \cdots + kw_n = k(w_1 + \cdots w_n) \leq k(v_1 + \cdots + v_m) = kv_1 + \cdots kv_m = c_{G'}(X', Y')$$

since $k > 0$. Therefore, the capacity of (X, Y) in G' is less than or equal to the capacity of (X', Y') in G'. Since (X', Y') is the min-cut of G', this implies (X, Y) is a min-cut of G', which contradicts our assumption that multiplying the edge capacities by k would not change the min-cut. By contradiction, this concludes the proof that the min-cut will not change after all edge capacities are multiplied by a positive number k.

3.

(15pts) You are given a flow network with unit-capacity edges. It consists of a directed graph $G = (V, E)$ with source s and sink t, and $c_e = 1$ for every edge e. You are also given a positive integer parameter k. The goal is to delete k edges so as to reduce the maximum s-t flow in G by as much as possible. In other words, you should find a subset of edges $F \subseteq E$ such that $|F| = k$ and the maximum s-t flow in the graph $G' = (V, E - F)$ is as small as possible. Give a polynomial-time algorithm to solve this problem and briefly explain its correctness.

Follow up: If the edges have more than unit capacity, will your algorithm produce the smallest possible max-flow value?.

Solution.
Note: Since $c_e = 1$ for all edges $e \in E$, we know that, if the maximum flow of G is $v(f)$, then the min-cut, (X, Y) has $v(f)$ edges from X to Y (since $c(X, Y) = v(f)$). Consider a different cut (X', Y') where $c_G(X', Y') > v(f)$. Then removing k edges from X' to Y' results in a capacity of $c_{G'}(X', Y') = c_G(X', Y') - k > v(f) - k$. However, if we remove k edges from X to Y, we have $c_{G'}(X, Y) = v(f) - k$. Since removing edges from X to Y results in a cut with smaller capacity than removing edges from X' to Y', we know that we should remove edges from X to Y in order to minimize the maximum flow in G'. We can use Edmond-Karp and BFS to find the min-cut in G, (X, Y), in $O(nm^2)$ time. We can add edges to F that go from X to Y one-by-one until $|F| = k$ or all edges from X to Y are in F. If all edges from X to Y are in F before $|F| = k$, we can arbitrarily add edges to F until $|F| = k$, because the maximum flow in G' will already be 0.
The described algorithm works as follows:

FindEdgesForRemoval(G, k)
Let G_f = the final residual graph from running Edmond-Karp on G with

```
    start node s and sink node t
    Run BFS on G_f with source node s, adding all reachable nodes to the set X
    Add all nodes in V − X to Y
    Let count = 0
    Let F = ∅
    For all edges (u, v) ∈ E
        if u ∈ X and v ∈ Y
            if count < k
                add (u, v) to F
                increment count by 1
            endIf
        endIf
    endFor
    if count < k
        For all edges (u, v) ∈ E
            if u ∉ X or v ∉ Y
                if count < k
                    add (u, v) to F
                    increment count by 1
                endIf
            endIf
        endFor
    endIf
    return F
endFindEdgesForRemoval
```

Time Complexity Analysis:
It takes $O(nm^2)$ time to run Edmond-Karp and store the final residual graph in G_f.
It takes $O(m+n)$ time to run BFS on G_f.
It takes $O(nlogn)$ time to create the X and Y sets.
It takes constant time to initialize *count* and F.
There are $O(n)$ iterations of the first for loop, each of which takes maximally $O(log^2m)$ time, for a total runtime of $O(nlog^2m)$ for the first for loop.
There are $O(n)$ iterations of the second for loop, each of which takes maximally $O(log^2m)$ time, for a total runtime of $O(nlog^2m)$ for the second for loop.
This results in a total runtime of $O(nm^2)+O(m+n)+O(nlogn)+O(1)+O(nlog^2m)+O(nlog^2m) = O(nm^2)$ for our algorithm. Thus, our algorithm has $O(nm^2)$ total runtime, so it is polynomial as required.

Follow Up:
This algorithm will not produce the smallest possible max-flow value if the edges have more than unit capacity. Our algorithm relies on the assumption that the cut (X, Y) with the least edges from X to Y is the min-cut. However, this assumption only holds when all edges have capacity 1. When edges have more than unit capacity, there could be one cut with only one edge from X to Y that has capacity 100, while another cut has 2 edges from X to Y, both with capacity 2. The min-cut in this case would be the cut with 2 edges from X to Y, showing how the assumption breaks down with edge capacities greater than 1. Thus, the assumption central to our algorithm breaks down when edges have capacities greater than 1, so our algorithm will not work for edges with more than unit capacity.

4.

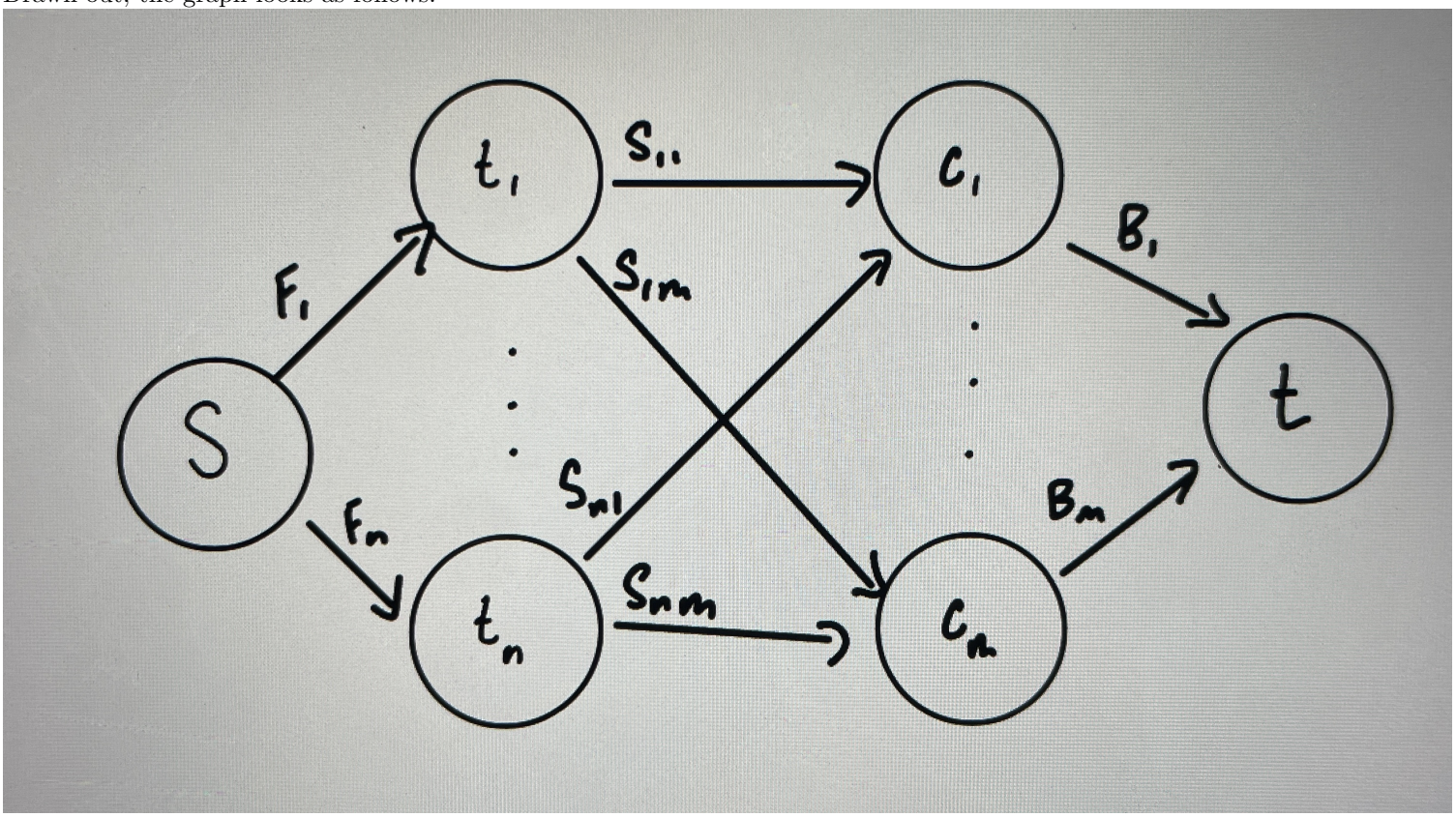
(20pts) A toursit group needs to convert their USD into various international currencies. There are n tourists $t_1, ..., t_n$ and m currencies $c_1, ..., c_m$. Each tourist t_i has F_i Dollars to convert. For each currency c_j, the

bank can convert at most B_j Dollars to c_j. Tourist t_i is willing to trade as much as S_{ij} of his Dollars for currency c_j. (For example, a tourist with 1000 dollars might be willing to convert up to 200 of his USD for Indian Rupees, up to 500 of his USD for Japanese Yen, and up to 300 of his USD for Euros). Assume that all tourists give their requests to the bank at the same time.

(a) Design an algorithm that the bank can use to satisfy all the requests (if it is possible). To do this, construct and draw a network flow graph, with appropriate source and sink nodes, and edge capacities.

(b) Prove your algorithm is correct by making a claim and proving it in both directions.

Solution. We will use a network flow graph, $G = (V, E)$, with source node s and sink node t. There will be a node for each t_i and a node for each c_j. That is, $V = \{s, t_1, ..., t_n, c_1, ..., c_m, t\}$.
There will be a directed edge from s to each t_i with capacity F_i. This ensures no tourist t_i converts more than F_i Dollars to different currencies. There will be a directed edge from each tourist t_i to each currency c_j with capacity S_{ij}. This ensures no tourist t_i converts more than S_{ij} Dollars to c_j. There will be a directed edge from each currency c_j to the sink node t with capacity B_j. This ensures the bank never converts more than B_j Dollars to currency c_j. That is, $E = \{(s, t_i, F_i), (t_i, c_j, S_{ij}), (c_j, t, B_j) | 1 \leq i \leq n, 1 \leq j \leq m\}$.
Drawn out, the graph looks as follows:



Once we construct the network flow graph, G, we can simply run Ford-Fulkerson on G. If all tourists can convert all F_i of their Dollars, then the flow through the graph should be $\sum_{i=1}^{n} F_i$. Therefore, we can determine if the Bank can satisfy all the requests by comparing the max-flow, $v(f)$, found by running Ford-Fulkerson on G to $\sum_{i=1}^{n} F_i$. If the values are equal, then the return value of Ford-Fulkerson, f, is the flow network that shows the banks how satisfy all requests. Thus, creating G essentially solves the problem for us, as we can just use the return value of Ford-Fulkerson to determine whether all requests can be satisfied and, if so, how to satisfy them.
Therefore, our algorithm works as follows:

canSatisfyRequests(t, c, F, B)

Construct $G = (V, E)$ as described above
Let f = return value of running Ford-Fulkerson on G with source node s
and sink node t
Let $v(f) = 0$
Let $needed = \sum_{i=1}^{n} F_i$
For all edges (s, v) in f
    add the flow through (s, v) to $v(f)$
endFor
if $v(f) \neq needed$ return false
return true
endCanSatisfyRequests

If the algorithm returns true, the flow network needed to satisfy all requests is stored in f. Otherwise, there is no way to satisfy all requests.

(b) **Proof of Correctness:**
Claim: The algorithm returns true $\iff$ all tourists can convert all F_i of their Dollars.
We must show that, if the algorithm returns true, all tourists can convert all F_i of their Dollars. Assume to the contrary that the algorithm returns true but some tourist cannot convert all F_i of his dollars. Since the algorithm returns true, we know $v(f) = needed = \sum_{i=1}^{n} F_i$. Therefore, the max flow from s to t in G is the sum of all F_i. Since the only edges exiting s in G are edges to t_i with capacity F_i, (s, t_i) must be saturated for all $1 \leq i \leq n$. Since the flow into each t_i must equal the flow exiting each t_i in f, this means that the flow through each t_i equals F_i for all $1 \leq i \leq n$. Thus, each tourist t_i can convert all F_i of their dollars, which contradicts our assumption that some tourist t_i cannot convert all F_i of his Dollars. This contradiction completes the proof that, if our algorithm returns true, all tourists can convert all F_i of their Dollars.
Now, we must show that, if all tourists can convert all F_i of their dollars, the algorithm will return true. If each tourist can convert all F_i of their dollars, then the max flow through each t_i will be (simultaneously) F_i for all $1 \leq i \leq n$. Since the flow into each t_i must equal the flow out of each t_i, this means there is F_i flow into each t_i in f for all $1 \leq i \leq n$. Since the only edges into t_i come from s, this means that there is an edge out of s with flow F_i in f for all $1 \leq i \leq n$. Since the max flow is the sum of the flows through edges exiting s in f, this means the max flow in G is $v(f) = \sum_{i=1}^{n} F_i$. Therefore, after our algorithm uses a for loop to calculate $v(f)$, $v(f)$ will always equal $\sum_{i=1}^{n} F_i$ if all tourists can convert all F_i of their dollars. Therefore, our algorithm will always return true after comparing $v(f)$ and $\sum_{i=1}^{n} F_i$.
This concludes the proof that our algorithm returns true $\iff$ all tourists can convert all F_i of their dollars.

5.

You have successfully computed a maximum s-t flow f for a network $G = (V; E)$ with integer edge capacities. Your boss now gives you another network G' that is identical to G except that the capacity of exactly one edge is decreased by one. You are also explicitly given the edge whose capacity was changed. Describe how you can compute a maximum flow for G' in $O(|V| + |E|)$ time.

Solution.
Note: We are using the representation of residual graphs that uses reverse edges, as discussed in lecture.
Suppose that the edge with decreased capacity is $(u, v) \in E$.
Suppose the final residual graph of G is G_f.
Suppose the max flow graph for G' is f'.
Suppose the final residual graph for G' is $G_{f'}$.
If (u, v) is the bottleneck of some s-t path(s) in f, then the flow through (u, v) in f must equal $c(u, v)$. Since the capacity of (u, v) decreases by 1 in G', this means the flow through (u, v) must equal $c(u, v) - 1$ in f'. Therefore, the flow out of v and the flow into u must decrease by 1 in f' compared to f. Therefore, in one

s-t path through (u, v), the flow through each edge on that path must decrease by 1 in f' compared to f to maintain conservation of flow. Therefore, if (u, v) is the bottleneck of some s-t path(s) in f, we can reduce the capacity of all edges along one such path in f by 1 to produce f'. To produce $G_{f'}$, we simply reduce the capacity of each edge in the reverse t-s path in G_f by 1 and increase the capacity of the forward edges in the s-t path in G_f by 1, excluding (u, v) itself.
If (u, v) is not the bottleneck any s-t path in f, then the residual capacity of (u, v) in G_f is positive, so we can reduce this value by 1, and the all s-t paths through (u, v) will still have the same flow. Thus, we can find $G_{f'}$ by simply reducing the residual capacity of (u, v) in G_f by 1 and not changing any other capacities. In this case, the max flow network for G' is the same as the max flow network for G, so $f' = f$.
Therefore, we just need to determine if (u, v) is the bottleneck of some s-t path(s) in f, which we can do by checking if (u, v) is in G_f. If it is not, we can run BFS on G_f twice, once from t to v and once from u to s, to find an s-t path through (u, v). Once we have found one such path, we can alter the edge capacities in G_f and f as described above to construct $G_{f'}$ and f'.
This algorithm works as follows:

alteredMaxFlow(f, G, G_f, (u, v))
    let $G_{f'} = G_f$
    let $f' = f$
    if $(u, v) \in G_f$
        reduce capacity of (u, v) in $G_{f'}$ by 1
        return f'
    endIf
    Run BFS on $G_{f'}$ with start node t until v is explored
    let $path = null$
    let $temp = v$
    while $predecessor(temp)! = null$
        add $(predecessor(temp), temp)$ to $path$
        let $temp = predecessor(temp)$
    endWhile
    Run BFS on $G_{f'}$ starting from u until s is explored
    let $temp = s$
    while $predecessor(temp)! = null$
        add $(predecessor(temp), temp)$ to $path$
        let $temp = predecessor(temp)$
    endWhile
    for all (x, y) in $path$
        decrease the capacity of (x, y) by 1 in $G_{f'}$
        if $x \neq u$ or $y \neq v$
            increase the capacity of (y, x) by 1 in $G_{f'}$
        endIf
        decrease the capacity of (y, x) by 1 in f'
    endFor
    return f'
endAlteredMaxFlow

Time Complexity Analysis:
It takes $O(|V| + |E|)$ time to initialize f' and $G_{f'}$.
It takes $O(log|E|)$ time to check if (u, v) a bottleneck by checking if it is in G_f.
It takes $O(|V| + |E|)$ time to run BFS on $G_{f'}$ from t to v and it takes $O(|V| + |E|)$ time to run BFS on $G_{f'}$ from u to s.
For each BFS call, it takes $O(|V|)$ time to update $path$.
It takes $O(|E|)$ time to update the edge capacities in $G_{f'}$ and f' after constructing $path$.
Thus, the algorithm has a total runtime of $O(|V|+|E|)+O(log|E|)+O(|V|+|E|)+O(|V|+|E|)+O(|V|)+$

$O(|V|) + O(|E|) = O(|V| + |E|)$, as required.

Assignment 9

Problem 1

In the network G below, the demand values are shown on vertices (supply value if negative). Lower bounds on flow and edge capacities are shown as (lower bound, capacity) for each edge. Determine if there is a feasible circulation in this graph. You need to show all your steps. (25 pt)

(a) Reduce the Feasible Circulation with Lower Bounds problem to a Feasible Circulation problem without lower bounds. (10 pt)

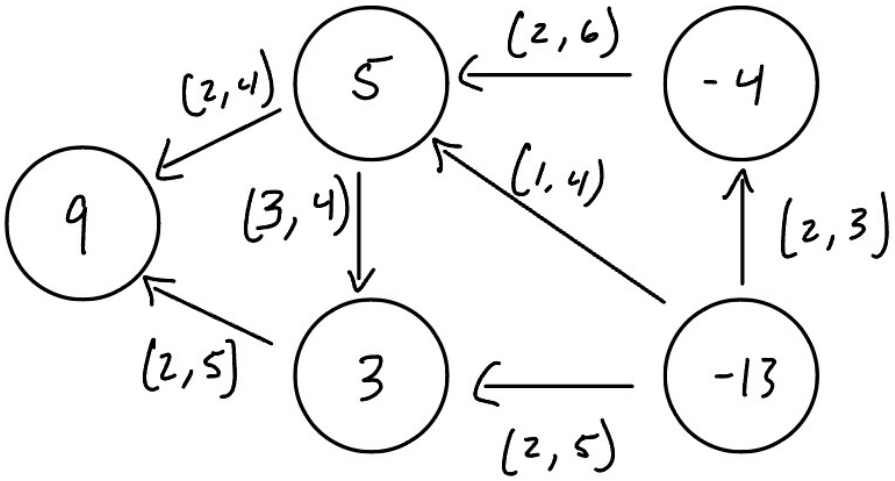
(b) Reduce the Feasible Circulation problem obtained in part (a) to a Maximum Flow problem in a Flow Network. (10 pt)

(c) Using the solution to the resulting Max Flow problem, explain whether there is a Feasible Circulation in G. (5 pt)

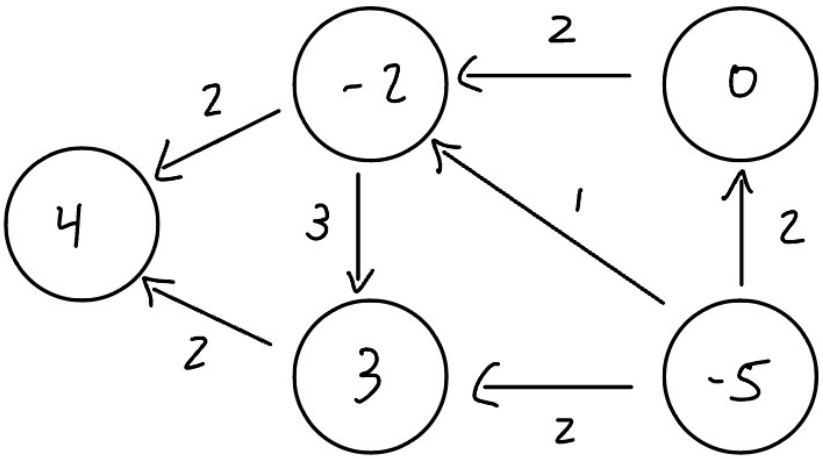
Solution. Note: We assume $G = (V, E)$.

(a) To reduce the Feasible Circulation with Lower Bounds problem to a Feasible Circulation problem without lower bounds, for all $e \in E$, we simply push flow equal to the lower bound of e through e. We can do so as follows:

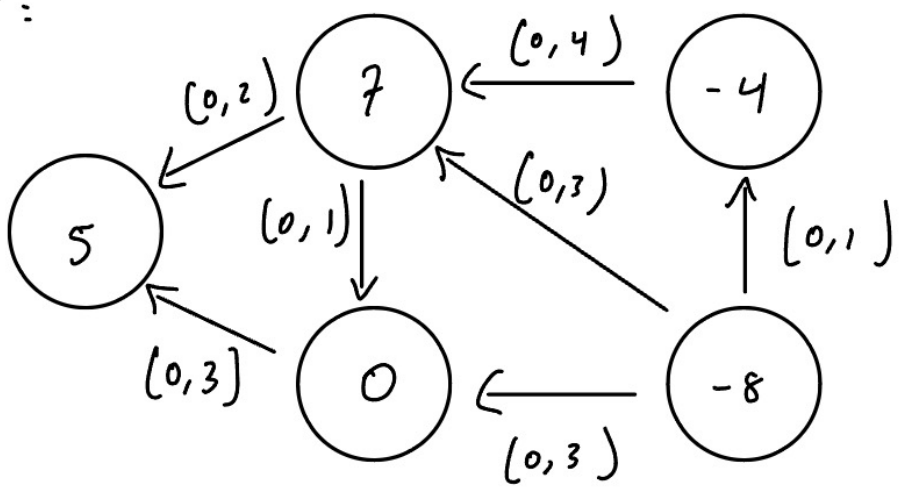
G_1:


5
(2, 6)
-4
(2, 4)
9
(3, 4)
(1, 4)
(2, 3)
(2, 5)
3
-13
(2, 5)

f_0:

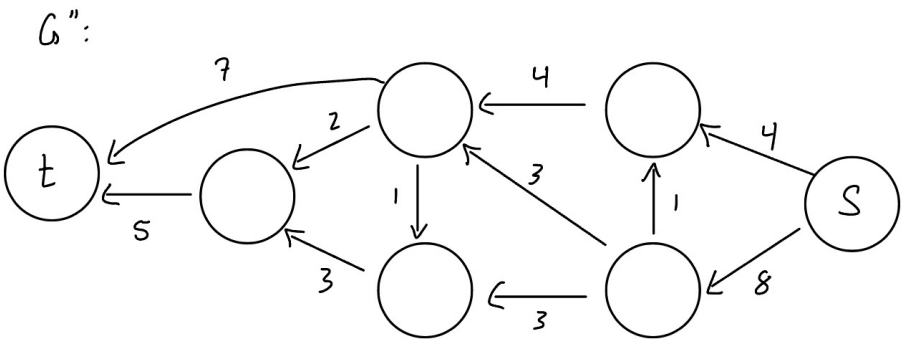

2
-2
0
2
4
3
1
2
2
3
-5
2

G':


(0, 4)
7
-4
(0, 2)
5
(0, 1)
(0, 3)
(0, 1)
(0, 3)
0
-8
(0, 3)

The graph $G' = (V', E')$ is now a Feasible Circulation problem in which all Lower Bounds equal 0, so we can treat it as a feasible circulation problem without Lower Bounds. There is a Feasible Circulation in $G' \iff$ there is a Feasible Circulation in G, so we have successfully reduced the problem to a Feasible Circulation problem without Lower Bounds.

(b) To reduce the Feasible Circulation problem without Lower Bounds from part (a) to a Maximum Flow problem in a Flow Network, we need to construct a valid Flow Network $G" = (V", E")$ from G'. To do this, we add a source node s and a sink node t. We then add a directed edge from s to all supply nodes from G' (all nodes with negative demand) and a directed edge from all demand nodes in G' (all nodes with positive demand) to t. We can do this as follows:



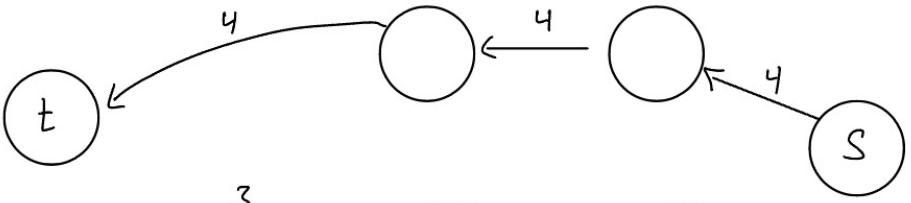
The graph $G"$ is now a valid flow network for which we can compute the Maximum Flow. There is a Feasible Circulation in $G \iff$ there is a Feasible Circulation in $G' \iff$ the value of the Max Flow in $G"$ is

$$v(f_1) = D_{G'} = \sum_{v \in V' : d_v > 0} d_v = 7 + 5 = 12$$

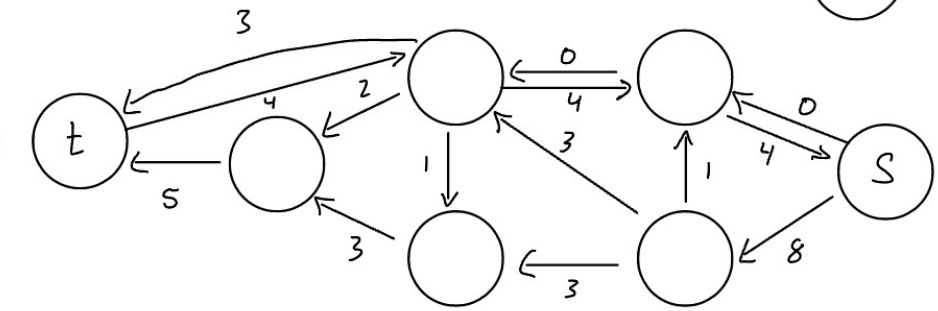
Thus, we have successfully reduced the problem to a Maximum Flow problem in a Flow Network.

(c) To find the value of the Maximum Flow $v(f_1)$ in $G"$, we run Ford-Fulkerson on $G"$ to produce:

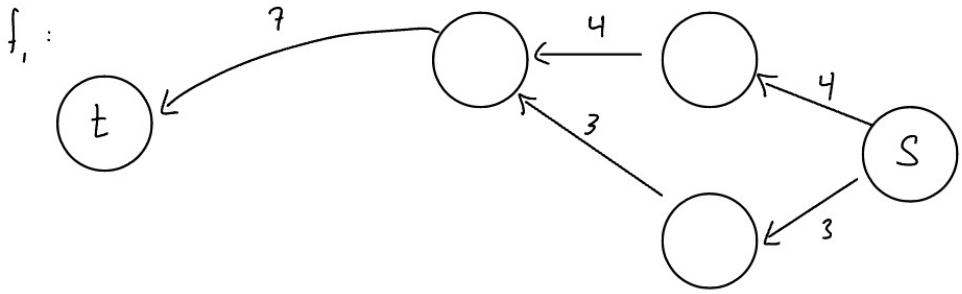
f_1 :



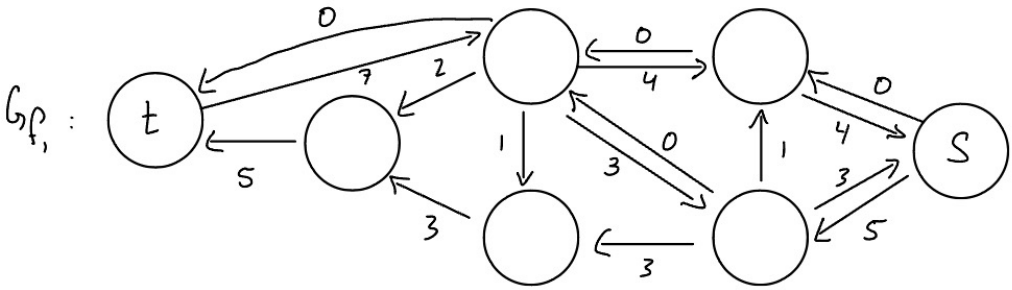
G_{f_1} :



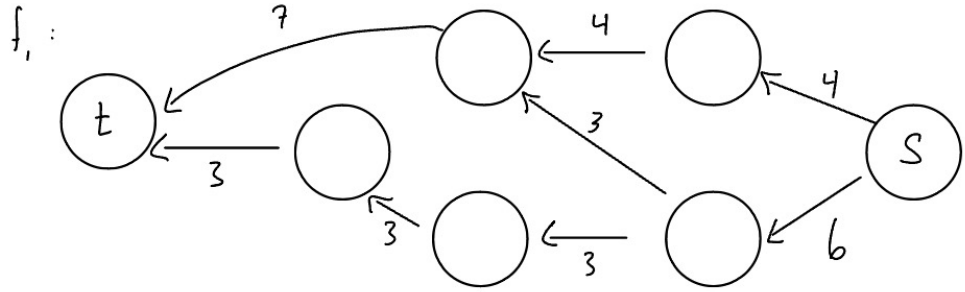
f_1 :



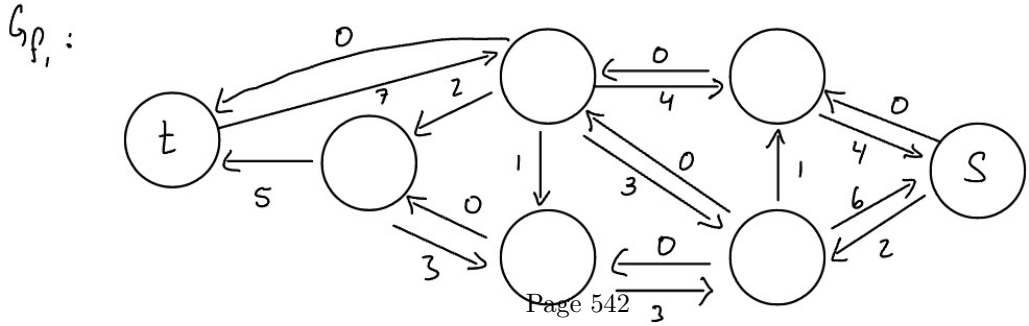
G_{f_1} :



f_1 :



G_{f_1} :



The sum of the flow exiting s in f_1 is $6 + 4 = 10$, so the value of the Max Flow in G" is $v(f_1) = 10$. Since

$$v(f_1) = 10 \neq 12 = D = \sum_{v \in V' : d_v > 0} d_v$$

we know that there is no Feasible Circulation in G', which implies there is no Feasible Circulation in G. We could also note that the value of the Min-Cut in G" is $v(f_1) = 3 + 3 + 4 = 10 \neq 12$ to conclude that there is no Feasible Circulation in either G' or G.

Thus, by reducing the Feasible Circulation with Lower Bounds problem into a Feasible Circulation without Lower Bounds problem and then a Maximum Flow in a Flow Network problem, we determined that there is no Feasible Circulation in G.

Problem 2

Solve Kleinberg and tardos, Chapter 7, Exercise 31. (25 pt)

Some of your friends are interning at the small high-tech company WebExodus. A running joke among the employees there is that the back room has less space devoted to high-end servers than it does to empty boxes of computer equipment, piled up in case something needs to be shipped back to the supplier for maintainence.

A few days ago, a large shipment of computer monitors arrived, each in its own large box; and since there are many different kinds of monitors in the shipment, the boxes do not all have the same dimensions. A bunch of people spent some time in the morning trying to figure out how to store all these things, realizing of course that less space would be taken up if some of the boxes could be nested inside others.

Suppose each box i is a rectangular parallelepiped with side lengths equal to (i_1, i_2, i_3); and suppose each side length is strictly between half a meter and one meter. Geometrically, you know what it means for one box to nest inside another: It's possible if you can rotate the smaller so that it fits inside the larger in each dimension. Formally, we can say that box i with dimensions (i_1, i_2, i_3) nests inside box j with dimensions (j_1, j_2, j_3) if there is a permutation a, b, c of the dimensions $\{1, 2, 3\}$ so that $i_a < j_1$, and $i_b < j_2$, and $i_c < j_3$. Of course, nesting is recursive: If i nests in j, and j nests in k, then by putting i inside j inside k, only box k is visible. We say that a *nesting arrangement* for a set of n boxes is a sequence of operations in which a box i is put inside another box j in which it nests; and if there were already boxes nested inside i, then these end up inside j as well. (Also notice the following: Since the side lengths of i are more than half a meter each, and since the side lengths of j are less than a meter each, box i will take up more than half of each dimension of j, and so after i is put inside j, nothing else can be put inside j.) We say that a box k is visible in a nesting arrangement if the sequence of operations does not result in its ever being put inside another box.

Here is the problem faced by the people at WebExodus: Since only the visible boxes are taking up any space, how should a nesting arrangement be chosen so as to minimize the *number* of visible boxes?
Give a polynomial-time algorithm to solve this problem.

Solution. Suppose there are n boxes in total, which we will call $b_1, \cdots, b_n$. Let $x =$ the number of visible boxes in some nesting arrangement. Let $y =$ the number of nested boxes in some nesting arrangement. Then

$$x = n - y$$

so we can minimize the number of visible boxes in a nesting arrangement by maximizing the number of nested boxes. We know that the we can use Edmonds Karp to find the value of maximum flow $v(f)$ in a flow network $G = (V, E)$, so we just need to construct G s.t. the value of maximum flow in G is $v(f) =$ the maximum number of nested boxes in any nesting arrangement. To do so, we can create two nodes for each

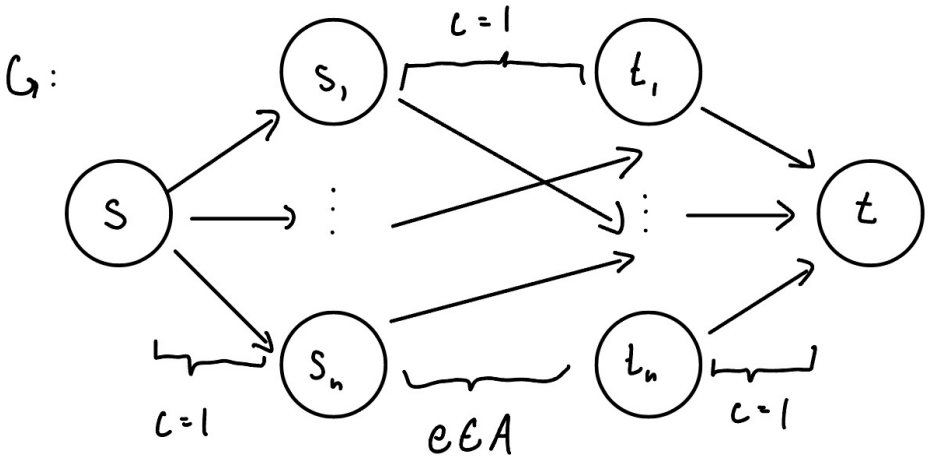
box b_i: a source node s_i and a sink node t_i. The source node s_i represents a box being nested inside box b_i. The sink node t_j represents the box b_j being nested inside another box. To satisfy the conditions of a flow network, we also create a supersource node s and a supersink node t. Thus, we have

$$V = \{s, s_1, \cdots, s_n, t_1, \cdots, t_n, t\}$$

For edges, we need a directed edge from the supersource node s to all source nodes s_i of capacity 1 to ensure that each path out no box b_i has more than one box nested directly inside of it (i.e. without being recursively nested). We also need a directed edge from all sink nodes t_i to the supersink node t of capacity 1 to ensure no box b_i is nested directly inside more than one box (i.e. without being recursively nested). Between source nodes s_i and sink nodes t_j s.t. $i \neq j$, we need a directed edge from s_i to t_j of capacity 1 $\iff$ box b_j can be nested inside box s_i. This ensures that the maximum flow in G represents a valid nesting arrangement. Thus, we have

$$E = \{(s, s_1, 1), \cdots, (s, s_n, 1), (t_1, t, 1), \cdots (t_n, t, 1), A\}$$

where $A = \{(s_i, t_j, 1) |$ box b_j can be nested inside of box $b_i\}$. Drawn out, the graph looks as follows:



Once we run Edmonds Karp to find the value of max flow $v(f)$ = the number of nested boxes in any nesting arrangement, we know that the minimum number of visible boxes is $x = n - v(f)$, so we can return this value.

In summary, our algorithm works as follows:

(i) Let n = the number of boxes

(ii) Construct $G = (V, E)$ as described above

(iii) Run Edmonds-Karp on G to find the value of max flow $v(f)$

(iv) Return $n - v(f)$

Proof of Correctness:
It suffices to show that the number of nested boxes in any nesting arrangement is y $\iff$ there is a valid flow f in G s.t. $v(f) = y$.
First, we will show that number of nested boxes in any nesting arrangement is y $\implies$ the value of the a valid flow f in G is $v(f) = y$:
If the number of nested boxes in any nesting arrangement is y, then we know that there are y pairs (b_i, b_j) s.t. box b_j can be nested in box b_i, and if (b_i, b_j) is a pair, then (b_i, b_k) and (b_k, b_j) are not pairs for all $1 \leq k \leq n$. Therefore, we can create a matching of size y between $\{s_1, \cdots, s_n\}$ and $\{t_1, \cdots, t_n\}$. Suppose Y = the set of such pairs, where $|Y| = y$. Then we can push 1 flow through each edge $(s_i, t_j, 1)$ s.t. $(b_i, b_j) \in Y$. Then, to satisfy conservation of flow, for all $(b_i, b_j) \in Y$, we push 1 flow from s to s_i and 1 flow from t_j to t. This

results in a valid flow f with y edge disjoint paths from s to t, each of which has a bottleneck of 1, which has value $v(f) = y$. Therefore, we have shown that, if the number of nested boxes in any nesting arrangement is y, there exists a valid flow f in G s.t. $v(f) = y$.
Next, we will show that the existence of a valid flow f in G of value $v(f) = y \implies$ there exists a nesting arrangement in which y boxes are nested.
If there is a valid flow f in G s.t. $v(f) = y$, then we must have y edge disjoint paths from s to t in G (since all edges have unit capacities). This means we can find y edges entering y distinct nodes t_j from y distinct nodes s_i s.t. box b_j can be nested inside box b_i. To construct a valid nesting arrangement s.t. the number of nested boxes is y, we simply nest box b_j inside box b_i for all $(s_i, t_j, 1)$ in f. Therefore, if there is a valid flow f in G of value $v(f) = y$, there exists a nesting arrangement in which exactly y boxes are nested.
This completes the proof that the number of nested boxes in any nesting arrangement is $y \iff$ there is a valid flow f in G s.t. $v(f) = y$. A direct corollary of this proof is that the value of max flow $v(f)$ in G is $y \iff y =$ the maximum number of nested boxes in any nesting arrangement. This corollary proves the correctness of our algorithm.

Time Complexity Analysis:
There are $2n + 2 = O(n)$ nodes in G, which take $O(n)$ time to create. There are n edges involving s, n edges involving t, and $\leq n^2$ edges between s_i and t_j, for a total of $n + n + n^2 = O(n^2)$ total edges. Since all edges e have capacity $c_e = 1$, it takes constant time to create each edge, so it takes $O(n^2)$ time to create them all. Therefore, it takes $O(n + n^2) = O(n^2)$ time to create G.
It takes $O(|V||E|^2)$ time to run Edmonds-Karp, at which point it takes constant time to compute the max flow value $v(f)$ and compute/return the minimum number of visible boxes in any nesting arrangement. Based on the previous calculations, in terms of n, Edmonds-Karp takes $O(n(n^2)^2) = O(n \cdot n^4) = O(n^5)$ time. Thus, the overall worst-case asymptotic complexity of our algorithm is $O(n^2) + O(n^5) = O(n^5)$, so our algorithm runs in polynomial time as required.

Problem 3

At a dinner party, there are n families $\{a_1, ..., a_n\}$ and m tables $\{b_1, ..., b_m\}$. The ith family a_i has g_i members and the jth table b_j has h_j seats. Everyone is interested in making new friends, and the dinner part planner wants to seat people such that no two members of the same family are seated in the same table. Design an algorithm that decides if there exists a seating assignment such that everyone is seated and no two members of the same family are seated at the same table. (25 pt)

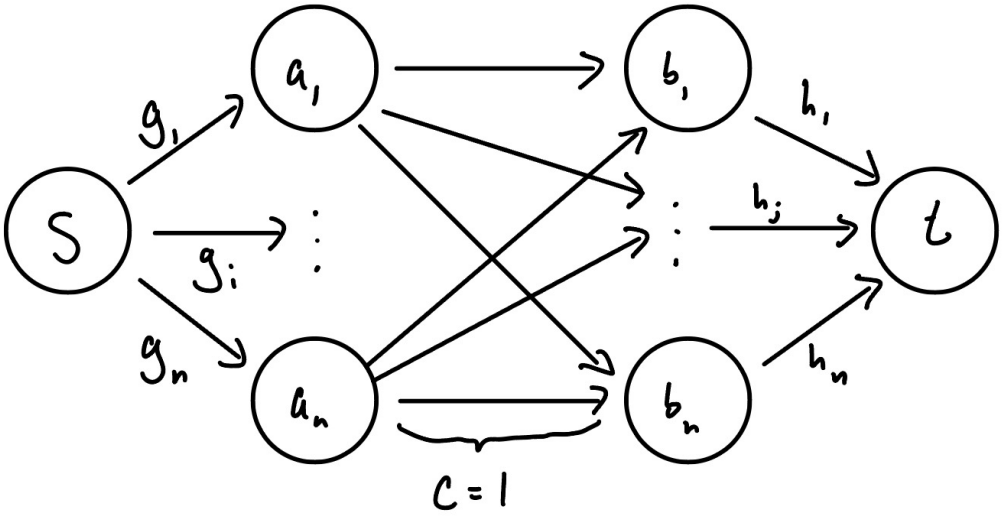
Solution.
We will reduce this problem to a Maximum Flow problem in a Flow Network. To construct our flow network $G = (V, E)$, we need to create a source node s and a sink node t. To represent our families, we will use n nodes $a_1, \cdots, a_n$. To represent our tables, we will use m nodes $b_1, \cdots, b_m$. Thus, we have

$$V = \{s, a_1, \cdots, a_n, b_1, \cdots, b_m\}$$

We need a directed edge from the source node s to each family node a_i of capacity g_i to represent the g_i family members in a_i. We need a directed edge of capacity 1 from each a_i to all b_j. This ensures that each family can send one member to any table, but no family can send more than one member to the same table. To ensure that no table b_j receives more than h_j assignments, we need a directed edge of capacity h_j from each table node b_j to the sink node t. Thus, we have

$$\begin{aligned} E = \quad & \{(s, a_1, g_1), \cdots, (s, a_n, g_n), \\ & (a_1, b_1, 1), \cdots, (a_1, b_m, 1), (a_2, b_1, 1), \cdots, (a_2, b_m, 1), \cdots (a_n, b_1, 1), \cdots, (a_n, b_m, 1), \\ & (b_1, t, h_1), \cdots (b_m, t, h_m)\} \end{aligned}$$

Drawn out, G looks as follows:



We can run Edmonds-Karp on G to find the value of the max flow, $v(f)$. Let

$$P = \sum_{i=1}^{n} g_i$$

Then we can return true if $v(f) = P$ and false otherwise.

In summary, our algorithm works as follows:

(i) Let $P = \sum_{i=1}^{n} g_i$

(ii) Construct $G = (V, E)$ as described above

(iii) Run Edmonds-Karp on G to find the value of the max flow $v(f)$

(iv) Return $v(f) == P$

Proof of Correctness:
It suffices to show that the value of max flow in G is $v(f) = P \iff$ all guests can be seated such that no two members of the same family are seated at the same table and no table b_j has more than h_j assignments.
First, we will show that value of max flow in G is $v(f) = P \implies$ all guests can be seated such that no two members of the same family are seated at the same table such that no table b_j has more than h_j assignments.
Since $v(f) = P$, we know all edges from s to a_i are saturated with g_i flow. By the conservation of flow, and since each edge exiting a_i has capacity 1, there must be g_i edges, each with flow 1, exiting each a_i. For each a_i, each of these g_i edges ends up at a unique b_i, so we know each member of each family can be assigned without any table receiving two members. Since f is a valid flow, we know that none of the b_j have more than h_j edges entering them. Therefore, if $v(f) = P$, we know that each member of each family can be assigned such that no two members of the same family are seated at the same table and no table b_j has more than h_j assignments.
Next, we will show that the ability to seat all guests such that no two members of the same family are seated at the same table and no table b_j has more than h_j assignments $\implies$ the value of the max flow in G is $v(f) = P$.
Since we can seat all guests in this way, we know we can direct g_i edges with flow 1 to unique tables b_j for each family a_i such that no table b_j has more than h_j assignments. Thus, we can direct edges with flow $\leq h_j$ from each table b_j to the sink node t. Since we know we have g_i flow out of each a_i for all $1 \leq i \leq n$,

we can direct g_i flow from the source node s to each family a_i to complete the construction of a valid flow. At this point, the flow out of s is

$$v(f) = \sum_{i=1}^{n} g_i = P$$

Also, since all edges out of s are saturated, we know that $v(f) = P$ is the value of the max flow in G. Thus, if we can seat all guests such that no two members of the same family are seated at the same table and no table b_j has more than h_j assignments, we know that the value of max flow in G is $v(f) = P$.
This completes the proof that the value of max flow in G is $v(f) = P \iff$ all guests can be seated such that no two members of the same family are seated at the same table and no table b_j has more than h_j assignments. The correctness of our algorithm follows from its return statement.

Time Complexity Analysis:
There are $n + m + 2 = O(n + m)$ nodes in G, which take a total of $O(n + m)$ time to create. There are n edges involving the source node s, m edges involving the sink node t, and nm edges between a family and a table. This results in a total of $m + n + mn = O(mn)$ edges, which take a total of $O(mn)$ time to create. Thus, it takes a total of $O(m + n) + O(mn) = O(mn)$ time to create G.
We can calculate P in $O(n)$ time directly. It takes constant time to return the answer once we run Edmonds-Karp and calculate the value of max flow $v(f)$ in G. It takes $O(|V||E|^2)$ time to run Edmonds-Karp. Since $|V| = O(m+n)$ and $|E| = O(mn)$, we know it takes $O((m+n)(mn)^2) = O((m+n)m^2n^2) = O(m^3n^2+n^3m^2)$ total runtime to perform Edmonds-Karp on G.
Thus, our algorithm has an overall worst-case asymptotic complexity of $O(m+n)+O(n)+O(m^3n^2+n^3m^2) = O(m^3n^2 + n^3m^2)$, so it is polynomial with respect to input size, as required.

Problem 4

Due to large-scale flooding in a region, paramedics have identified a set of n injured people distributed across the region who need to be rushed to hospitals. There are k hospitals in the region, and each of the n people needs to be brought to a hospital that is within a half-hour's drive to their current location. (So different patients will be able to be served by different hospitals depending upon the patients' locations.) However, overloading one hospital with too many patients at the same time is undesirable, so we would like to distribute the patients as evenly as possible across all the hospitals. So the paramedics (or a centralized service advising the paramedics) would like to work out whether they can choose a hospital for each of the injured people in such a way that each hospital receives at most $(\frac{n}{k} + 1)$ patients. (25 pt)

(a) Describe a procedure that takes the given information about the patients' locations (hence specifying which hospital each patient could go to) and determines whether a balanced allocation of patients is possible (i.e. each hospital receives at most $(\frac{n}{k} + 1)$ patients). (11 pt)

(b) Provide proof of correctness for your procedure. (10 pt)

(c) What is the asymptotic running time of your procedure (in terms of n and k)? (4 pt)

Solution.

(a) We will reduce this problem to a Maximum Flow problem in a Flow Network. To construct our flow network $G = (V, E)$, we need to create a source node s and a sink node t. We also need to create a node for each patient $p_1, ..., p_n$ and for each hospital $h_1, ..., h_k$. Thus, we have

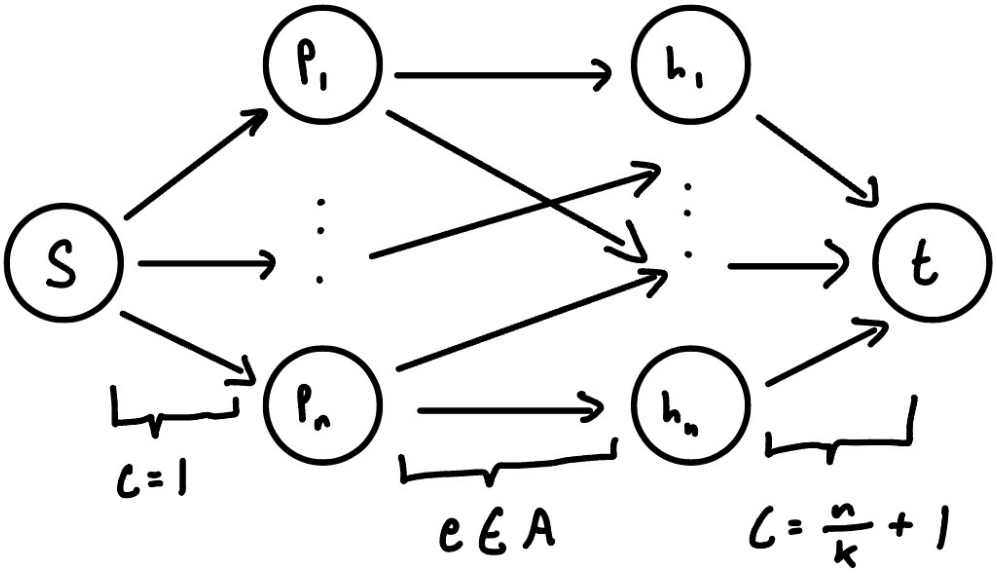
$$V = \{s, p_1, ..., p_n, h_1, ..., h_k, t\}$$

We need a directed edge to connect our source node s each of our patients p_i with a capacity of 1 to indicate that each p_i represents only 1 patients. We need a directed edge to connect each patient node p_i to each hospital node h_j s.t. h_j is within a 30 minute drive from p_i. Each of these edges should also have capacity 1 to indicate that each patient p_i represents a single patient. Since each hospital can hold

a maximum of $\frac{n}{k}+1$ patients, we need a directed edge from each hospital h_k to the sink node t with capacity $\frac{n}{k}+1$. Thus, we have

$$E = \{(s, p_1, 1), ..., (s, p_n, 1), (h_1, t, \frac{n}{k}+1), ..., (h_k, t, \frac{n}{k}+1), A\}$$

where
$A = \{(p_i, h_j, 1) |$ hospital h_j is within a 30 minute drive from patient $p_i\}$. Drawn out, the graph G looks as follows:



Once we perform Edmonds-Karp on G to find the value of maximum flow in G, $v(f)$, we can return true if $v(f) = n$ and false otherwise.

In summary, our algorithm works as follows:

(i) Let $n =$ the number of patients

(ii) Construct $G = (V, E)$ as described above

(iii) Run Edmonds-Karp to find the value of max flow $v(f)$ in G

(iv) Return $v(f) == n$

(b) **Proof of Correctness:**
It suffices to show that the value of max flow in G is $v(f) = n \iff$ a balanced allocation of patients is possible.
First, we will show that the value of max flow in G is $v(f) = n \implies$ a balanced allocation of patients is possible.
Since $v(f) = n$, we know all edges exiting s are saturated. By conservation of flow, we know there is an edge exiting each p_i of capacity 1. Also by conservation of flow, we know the total number of edges entering each h_j is $\leq \frac{n}{k}+1$. Thus, we can assign each patient p_i to the hospital h_j corresponding to the edge exiting p_i in f. This is a balanced allocation of patients. Thus, we have shown that $v(f) = n \implies$ a balanced allocation of patients is possible.
Next, we will show that the possibility of a balanced allocation $\implies$ the value of the max flow in G is $v(f) = n$.

Since a balanced allocation is possible, we can assign each of the n patients to one of the k hospitals such that no hospital receives $> \frac{n}{k} + 1$ patients. To construct a valid flow f in G, we can push 1 flow from each p_i to the hospital h_j that it is assigned to in the balanced allocation. We can then push flow equal to the number of patients assigned to each hospital h_j from h_j to t. To satisfy conservation of flow and complete the construction of f, we can push 1 flow from s to each patient p_i (since each p_i has 1 flow exiting it). At this point, f is a valid flow, and all n of the edges exiting the source node s are saturated with flow 1, so the value of flow is $v(f) = n$. Since the sum of the capacities of all edges exiting s in G is n, the max flow of any valid flow in G is always $\leq n$. Thus, since $v(f) = n$, we know f is a max flow. Thus, we have shown that, if a balanced allocation is possible, the value of the max flow in G is $v(f) = n$.
This completes the proof that the value of max flow in G is $v(f) = n \iff$ a balanced allocation of patients is possible. The correctness of our algorithm follows from its return statement.

(c) There are $n + k + 2 = O(n + k)$ nodes, which take a total of $O(n + k)$ runtime to create. There are n edges involving s, k edges involving t, and $O(nk)$ edges involving one patient p_i and one hospital h_j, for a total of $k + n + O(nk) = O(nk)$ edges. These take a total of $O(nk)$ runtime to create. Thus, the creation of G takes a total of $O(nk) + O(n + k) = O(nk)$ runtime.
Edmonds-Karp takes $O(|V||E|^2)$ runtime. Since $|V| = O(n + k)$ and $|E| = O(nk)$, we know running Edmonds-Karp on our algorithm takes

$$O((n + k)(nk)^2) = O((n + k)n^2k^2) = O(n^3k^2 + k^3n^2)$$

runtime. It takes constant time to compute and return the solution once we find $v(f)$ using Edmonds-Karp.
Thus, the worst-case asymptotic complexity of our algorithm is $O(n^3k^2+k^3n^2)+O(nk) = O(n^3k^2+k^3n^2)$ runtime, so it is polynomial with respect to input size, as required.

Assignment 10

Problem 1 (25pts)

Consider the partial satisfiability problem, denoted as 3-Sat(α). We are given a collection of k clauses, each of which contains exactly three literals, and we are asked to determine whether there is an assignment of true/false values to the literals such that at least αk clauses will be true. Note that 3-Sat(1) is exactly the 3-SAT problem from lecture.

Prove that 3-Sat($\frac{15}{16}$) is **NP**-complete.

Hint: If x, y, and z are literals, there are eight possible clauses containing them: $(x \vee y \vee z)$, $(!x \vee y \vee z)$, $(x \vee !y \vee z)$, $(x \vee y \vee !z)$, $(!x \vee !y \vee z)$, $(!x \vee y \vee !z)$, $(x \vee !y \vee !z)$, $(!x \vee !y \vee !z)$.

Proof. First, we must show that $3 - Sat(\frac{15}{16})$ is NP. In this case, the certificate t is an assignment of boolean values to all the literals, which is polynomial in length with respect to the number of clauses. To check if a certificate t is a valid assignment, we can just count how many clauses t satisfies and return true if $t \geq k\frac{15}{16}$, false otherwise. It takes constant time to evaluate each clause, so it takes polynomial time to certify the certificate t. Thus, $3 - Sat(\frac{15}{16})$ has both polynomial length certificate and a polynomial time certifier, so it has an efficient certification, so $3 - Sat(\frac{15}{16})$ is NP.
Next, we must reduce a known *NP-Complete* problem to $3 - Sat(\frac{15}{16})$. We will reduce $3 - Sat(1)$, which we know from lecture is *NP-Complete*. Consider any $3 - Sat(1)$ instance with k clauses. Consider any clause $c = (x_i \vee x_j \vee x_k)$. Following the hint, the eight disjunctions of the literals x_i, x_j, x_k and their complements are $(x_i \vee x_j \vee x_k)$, $(x_i \vee x_j \vee !x_k)$, $(x_i \vee !x_j \vee x_k)$, $(x_i \vee !x_j \vee !x_k)$, $(!x_i \vee x_j \vee x_k)$, $(!x_i \vee x_j \vee !x_k)$, $(!x_i \vee !x_j \vee x_k)$, and $(!x_i \vee !x_j \vee !x_k)$. For any truth assignment $(x_i = a, x_j = b, x_k = c)$, where $a, b, c \in \{$ true, false $\}$ the

only disjunction that evaluates to false is $(!a \vee !b \vee !c)$. Therefore, for any truth assignment to the literals (x_i, x_j, x_k), exactly 7 of the the eight disjunctions evaluate to true. We want to modify the $3-Sat(1)$ instance such that using our $3-Sat(\frac{15}{16})$ blackbox on the modified instance returns true $\iff$ we can satisfy the original $3-Sat(1)$ instance.Thus, for each clause $c = (x_i \vee x_j \vee x_k)$ in the $3-Sat(1)$ instance, we can create 16 new clauses. 9 of these clauses will be duplicates of c. The other 7 clauses will be the remaining disjunctions of x_i, x_j, x_k, namely $(x_i \vee x_j \vee !x_k)$, $(x_i \vee !x_j \vee x_k)$, $(x_i \vee !x_j \vee !x_k)$, $(!x_i \vee x_j \vee x_k)$, $(!x_i \vee x_j \vee !x_k)$, $(!x_i \vee !x_j \vee x_k)$, and $(!x_i \vee !x_j \vee !x_k)$. We can then use our $3-Sat(\frac{15}{16})$ blackbox with $16k$ clauses to determine if there is an assignment that satisfies $\geq \frac{15}{16} \cdot 16k = 15k$ of the clauses. Note that we are creating $16 = O(1)$ new clauses, each in $O(1)$ time, for each of the k clauses in the initial $3-Sat(1)$ instance, so it takes $O(k)$ time to create the modified $3-Sat(\frac{15}{16})$ instance. Thus, we have reduced in polynomial time the $3-Sat(1)$ problem to the $3-Sat(\frac{15}{16})$ problem. To prove our reduction is valid, we must now prove that there is a satisfying assignment in the modified $3-Sat(\frac{15}{16})$ instance $\iff$ there is a satisfying assignment in the initial $3-Sat(1)$ instance.

Consider any clause c from the initial $3-Sat(1)$ instance. If $c = (x_i \vee x_j \vee x_k)$ evaluates to true, then all 9 copies of c in the modified $3-Sat(\frac{15}{16})$ instance evaluate to true, and 6 of the 7 remaining disjunctions of x_i, x_j, x_k and their complements evaluate to true. The last part follows since only $(!x_i \vee !x_j \vee !x_k)$ would evaluate to false. Thus, if c evaluates to true in some truth assignment, then exactly $9 + 6 = 15$ of the 16 clauses in the modified $3-Sat(\frac{15}{16})$ instance evaluate to true. Similarly, if if c evaluates to false, then all 9 copies of c in the modified instance evaluate to false, while the remaining 7 evaluate to true. In this case, exactly 7 out of the 16 clauses corresponding to c evaluate to true. This also implies that the number of clauses in the modified instance that evaluate to true should always be between $\frac{7}{16} \cdot 16k = 7k$ and $\frac{15}{16} \cdot 16k = 15k$.

If there is a satisfying assignment in the modified $3-Sat(\frac{15}{16})$ instance, then we know at least $\frac{15}{16}k$ clauses evaluate to true under that assignment. Since maximally $\frac{15}{16}k$ clauses in the modified instance evaluate to true under any assignment, we know exactly $\frac{15}{16}k$ clauses evaluate to true. If any clause c' from the initial instance evaluate to false, then we would have less than $\frac{15}{16}k$ clauses evaluating to true in the modified instance. This follows from only 7 of the 16 clauses corresponding to c' would evaluate to true under any truth assignment in the modified instance. Thus, if there is an assignment satisfying the modified $3-Sat(\frac{15}{16})$ instance, that assignment also satisfies the initial $3-Sat(1)$ instance.

Similarly, if there is a satisfying assignment for the initial $3-Sat(1)$ instance, then we know, for each clause c in the initial instance, that assignment causes 15 out of the 16 clauses corresponding to c to evaluate to true in the modified $3-Sat(\frac{15}{16})$ instance. Since this is true for all k clauses independently, we know that exactly $15k$ of the $16k$ clauses in the modified instance evaluate to true under the satisfying assignment for the initial instance. Thus, if there is an assignment satisfying the initial $3-Sat(1)$ instance with k clauses, it also satisfies the modified $3-Sat(\frac{15}{16})$ with $16k$ clauses.

This completes the proof that $3-Sat(\frac{15}{16})$ is *NP-Complete.* □

Problem 2

[25 pts.] Consider a graph $G = (V, E)$ and two integers k, m.

2a

A **k-clique** is a subset of nodes $u_i \in G$, $i = 1, ..., k$ such that there is an edge connecting each pair of distinct vertices u_i, u_j. In other words, the **k-clique** is a complete sub-graph of G. Prove that finding a clique of size k is NP-Complete. [15 pts.]

2b

The **Dense Subgraph** problem is to find a subset V' of V, whose size is at most k and is connected by at least m edges. Prove that the **Dense Subgraph** problem is NP-Complete. [10 pts.]

Solution.

(a)

Proof. First, we must show that finding a clique of size k is NP.

In this case the certificate t is a subset $U \subseteq G$. The subset must have size k, so it is polynomial length with respect to the input, and it must be complete. We can check if U has size k in $O(1)$ time. To check if U is complete, we must make sure that all nodes $u \in U$ have exactly $k-1$ edges between other nodes in U. This takes $O(k-1) = O(k)$ time for each of the k nodes in U, for a total of $O(k^2)$ time. Thus, it takes a total of $O(k^2)$ time to check if a subset of nodes U represents a valid clique of size k, so finding a clique of size k has a polynomial time certifier. Thus, finding a **k-clique** has an efficient certification, so it is NP.
Next, we must reduce a known *NP-Complete* problem to finding a clique of size k in a graph G. We can reduce the $3-SAT$ problem, which we know from lecture is *NP-Complete*. Consider a $3-SAT$ instance with k clauses. We will construct $G = (V, E)$ based on these clauses. For each literal x_i in each clause c_j, we will create a vertex v_{x_i,c_j}. We will create edges between all vertices v_{x_i,c_j}, v_{x_k,c_l} s.t. $x_i \neq !x_k$ and $j \neq l$. Then we will call our blackbox to see if we can find a **k-clique** in G. Note that there are $O(3k) = O(k)$ vertices, which take $O(k)$ time to create. There are $O(k-3) = O(k)$ edges for each of the $O(k)$ vertices, which take a total of $O(k^2)$ time. Thus, it takes polynomial time to create G, so we have reduced in polynomial time the $3-SAT$ problem to a finding a **k-clique** in a graph G. To prove our reduction is valid, we must now show that we can find a **k-clique** $\iff$ the $3-SAT$ instance is satisfiable.
If we can find a **k-clique** in G, then we know there is a complete subgraph with k vertices. Since no edges exist between literals in the same clause, we know these k vertices represent literals from k distinct clauses. Also, since no edges exist between a literal and its complement, we know we can set all literals corresponding to vertices in the **k-clique** to true without creating any contradictions. At this point, we have at least one literal from each of the k clauses in the $3-SAT$ instance set to true without contradictions, so we know all clauses evaluate to true. We can set all literals that do not appear in any vertices of the **k-clique** to false, and we have satisfied the $3-SAT$ instance.
If we have a satisfying assignment of literals for the $3-SAT$ instance, then we know at least one literal from each of the k clauses is set to true without contradictions. We can select the vertices corresponding to one such literal from each of these clauses and all the edges between them to be part of our **k-clique**. We know none of the corresponding literals are complements of each other because the satisfying assignment does not produce contradictions. Since there are edges between all vertices that are not part of the same clause, and since none of the selected vertices represent complements of each other, we know there are edges between each distinct pair of the k selected vertices in G. Therefore, we know we have formed a valid **k-clique**.
This completes the proof that finding a **k-clique** is *NP-Complete*

□

(b)

Proof. First, we must show that the **Dense Subgraph** problem is NP. In this case, the certificate t is a subset V' of V, which is polynomial length with respect to the graph G. The certifier must check if $|V'| \leq k$, which can be done in $O(1)$ time. It must also check if there are at least m edges connecting the vertices in V'. To do this, we can count all distinct edges between vertices in V' and compare the count to m. This can be done in $O(n^2)$ time, where n is the number of vertices in G, as each vertex v has maximally $n-1$ edges to distinct nodes. Thus, the **Dense Subgraph** problem has a polynomial length certificate and a polynomial time certifier, so it has an efficient certification, so it is NP.
Next, we must reduce a known *NP-Complete* problem to the **Dense Subgraph** problem. We will reduce the **k-clique** problem, which we know from part (a) is *NP-Complete*. Note that, in a **k-clique**, since each vertex has $k-1$ edges, but an edge from u to v is the same as an edge from v to u (assuming an undirected graph), there are exactly $\frac{k(k-1)}{2}$ edges in a **k-clique**. By definition, there are k vertices in a **k-clique**. Since a **k-clique** is a complete subgraph, we know that any collection of k vertices can have at most $\frac{k(k-1)}{2}$ edges connecting them.
Thus, to determine if there exists a **k-clique** in G, we can let $m = \frac{k(k-1)}{2}$ and use our blackbox to determine if there is a subset V' of V whose size is at most k and is connected by at least $m = \frac{k(k-1)}{2}$ edges. It takes $O(1)$ time to calculate the value of m, so we have reduced in polynomial time the **k-clique** problem to a **Dense Subgraph** problem with $k = k$, $m = \frac{k(k-1)}{2}$. To prove our reduction is valid, we must show that there is a **k-clique** in $G = (V, E)$ $\iff$ there is a subset V' of V of size at most k connected by at least

$\frac{k(k-1)}{2}$ edges.
If there exists a valid **k-clique**, then there exists a subgraph of G with k vertices and $\frac{k(k-1)}{2}$ edges. Thus, the **k-clique** itself forms subset V' of V of size at most k connected by at least $\frac{k(k-1)}{2}$ edges. This subset is our valid solution to the **Dense Subgraph** problem.
If there exists a subset V' of V of at most k vertices connected by at least $\frac{k(k-1)}{2}$ edges, then, since any subset of $\leq k-1$ vertices has at most $\frac{(k-1)(k-2)}{2}$ edges, we know that V' has exactly k vertices. Moreover, since any subset of k vertices is connected by at most $\frac{k(k-1)}{2}$ edges, we know V' is connected by exactly $\frac{k(k-1)}{2}$ edges. Thus, V' and the edges connecting distinct vertices in V' form a complete subgraph of size k. This is our valid **k-clique**.
This concludes the proof that the **Dense Subgraph** problem is *NP-Complete*. □

Problem 3 (25 pts)

Consider a modified SAT problem, SAT', in which, given a CNF formula having m clauses and n variables $x_1, x_2, ..., x_n$, the output is YES if there is an assignment to the variables such that exactly $m-2$ clauses are satisfied, and NO otherwise. Prove that SAT' is *NP-Complete*.

Proof. First, we must show that SAT' is *NP*. In this case, the certificate t is an assignment of true and false values to the n variables in the SAT' instance. This is polynomial in length with respect to n. The certifier must count how many of the m clauses are satisfied by the truth assignment. If this count equals $m-2$, the certifier should return true. Otherwise, the certifier should return false. It takes $O(n)$ to evaluate each of the m clauses under the truth assignment t, for a total of $O(mn)$ time. Thus, SAT' has both a polynomial length certificate and a polynomial time certifier, so it has an efficient certificate, so SAT' is *NP*.
Next, we must show that we can reduce a known *NP-Complete* problem to a SAT' problem in polynomial time. We will reduce SAT, which we know from discussion is *NP-Complete*. We want to produce a SAT' instance from a SAT instance such that the SAT' instance outputs YES $\iff$ the initial SAT instance is satisfiable. Consider a SAT instance with m clauses and n variables. To create our SAT' instance, we can add two new variables, y and z, and four new clauses, each which consist of one literal: (y), (!y), (z), (!z)
Thus, if the initial SAT instance had a logic formula of

$$c_1 \wedge c_2 \wedge ... \wedge c_m$$

then the modified SAT' instance has a logic formula of

$$c_1 \wedge c_2 \wedge ... \wedge c_m \wedge y \wedge !y \wedge z \wedge !z$$

Thus, SAT' has

$$m' = m + 4$$

clauses and

$$n' = n + 2$$

variables. In any truth assignment, exactly two of (y), (!y), (z), and (!z) are true, and the other two are false. Therefore, in any truth assignment, at most $m' - 2 = m + 2$ clauses are true in the SAT' instance. Thus, to determine if there is a truth assignment that satisfies all m clauses from the initial SAT instance, we can use our SAT' blackbox to determine if there is a truth assignment that satisfies exactly $m' - 2$ clauses in the modified SAT' instance. It takes constant time to to add the additional variables and clauses to create the modified SAT' instance, so we have reduced in polynomial time a SAT problem with m clauses and n variables to a SAT' problem with m' clauses and n' variables. To prove our reduction is correct, we must show that there exists a truth assignment that satisfies all m clauses from the initial SAT instance $\iff$ there exists a truth assignment that satisfies exactly $m' - 2 = m + 2$ clauses from the modified SAT' instance.
If there exists a truth assignment that satisfies all m clauses from the initial SAT instance, then since all m

clauses are also in the modified SAT' instance, we know at least m clauses are satisfied in the SAT' instance. Furthermore, since exactly two of (y), (!y), (z), and (!z) are true under any truth assignment, the satisfying truth assignment for the SAT instance must satisfy exactly $m + 2 = m' - 2$ clauses in the modified SAT' instance (assuming we modify it to assign values arbitrarily to y and z), so the truth assignment for the SAT instance also satisfies the SAT' instance for any y, z.
If there exists a truth assignment that satisfies exactly $m' - 2 = m + 2$ clauses in the modified SAT' instance, then since there are only m' clauses, and two of (y), (!y), (z), and (!z) must always evaluate to false, the truth assignment must satisfy exactly m clauses excluding (y), (!y), (z), and (!z). The only other m clauses in the SAT' instance are the m clauses from the initial SAT instance, so we know all m clauses from the initial SAT instance must evaluate to true under this truth assignment. Thus, the truth assignment that satisfies $m' - 2 = m + 2$ clauses in the SAT' instance also satisfies all m clauses from the SAT instance.
This completes the proof that SAT' is *NP-Complete*.

□

Problem 4 (25 pts)

Show that Vertex Cover is still *NP-Complete* even when all vertices in the graph are restricted to have even degree.

Proof. First, we must show that Vertex Cover with even degrees is *NP*. In this case, the certificate t is a subset V' of V, where $G = (V, E)$, so it has polynomial length with respect to input size. For each edge, the certifier ensure that at least one of the two points endpoints is an element of V'. This takes constant time for each of $m = |E|$ edges, so the certifier takes $O(m)$ time. Thus, Vertex Cover with even degrees has both a polynomial length certificate and polynomial time certifier, so it has efficient certification, so it is *NP*.
Next, we must reduce a known *NP-Complete* problem to Vertex Cover with even degrees. We will reduce Vertex Cover, which we know from lecture is *NP-Complete*. We want to modify the graph $G = (V, E)$ such that G has a vertex cover of size $k \iff$ the modified graph $G' = (V', E')$ has a vertex cover with even degrees of size $k + 2$. Since all degrees must be even, we can create a new vertex u in G', and connect to u edges from all vertices in V with odd degrees. This will make all nodes from V have even degrees. It will make u have degree equal to the number of nodes with odd degrees in V.

Claim: the number of nodes with odd degrees in V is even.

Proof. The sum of the degrees of all vertices in V is the number of endpoints of edges in G. Since each edge has two endpoints, and $m = |E|$ is an integer, the sum of the degrees of all vertices is equal to $2m = 2|E|$, which is even. Suppose the sum of the even-degree nodes in V is s. Then s is a sum of even numbers, so it is also even. The sum of the odd-degree nodes is $2m - s$, which is also even since it is the difference between two even numbers. For the sum of odd numbers to be even, there must be an even number of them added together. Therefore, since the sum of the degrees of the odd-degree vertices is even, we know there are an even number of nodes with odd-degrees. □

Since u has degree equal to the number of nodes with odd degrees in V, we know u has even degree. To ensure that a Vertex Cover in G' includes two removable nodes, we add two mode vertices to V', v and w, which are only attached to u and to each other. Note that v and w both have the even degree of 2. Now, if a vertex cover doesn't have u, it must have both v and w to cover the cycle between the three vertices. Since v and w are not connected to any nodes or edges from G, we can switch one of them with u, and we still have a vertex cover. Therefore, if we can produce a Vertex Cover in G' of size k, then we can produce a vertex cover in G' of size k that includes v. Since we removed nothing from G to produce G', this vertex cover should also act as vertex cover for G, with appropriate vertices removed. Thus, once we add u, v, w and corresponding edges to G to form G', we can determine if there is a Vertex Cover in G of size k by using our Vertex Cover with even-degrees blackbox to determine if there is a Vertex Cover in G' of size $k + 2$. It

takes $O(|V|) = O(n)$ time to add the edges from odd-degree nodes in G to u. It takes constant time to add v, w, and their edges. Thus, we have reduced in polynomial time a Vertex Cover problem to a Vertex Cover with even degrees problem. To prove our reduction is valid, we must show that there is a Vertex Cover with even degrees of size $k+2$ in $G' \iff$ there is a Vertex Cover of size k in G.
If we have a Vertex Cover V_G of size k in G, then the only edges in G' that might not be connected are edges involving u and the edge between v and w. Thus, we can add u and either of v and w to V_G to create a Vertex Cover $V_{G'}$ in G'. Since V_G has size k, and we add two vertices to V_G to create $V_{G'}$, we know $V_{G'}$ is a Vertex Cover of size $k+2$ in G'.
If we have a Vertex Cover $V_{G'}$ of size $k+2$ in G', then we know we must have at least two vertices among u, v, w in $V_{G'}$ in order to cover all edges in the cycle between these three vertices. Since none of these vertices cover any edges in G, we can remove two of them from $V_{G'}$ to create a Vertex Cover V_G in G. Since $V_{G'}$ has size $k+2$, and we removed two nodes to create V_G, we know V_G is a Vertex Cover in G of size k. One technicality is if all three of u, v, and w are in $V_{G'}$. After removing u and v, we have w left in our V_G, which doesn't make sense since w is not in G. However, since w covers none of the edges in G, we can just pick any vertex from G that is not yet in V_G to replace w. This results in a Vertex Cover V_G in G of size k, where all vertices are guaranteed to be in G itself.
Thus, there exists a Vertex Cover of size $k+2$ in $G' \iff$ there exists a Vertex Cover of size k in G. This concludes the proof that Vertex Cover is still *NP-Complete*, even when all vertices in the graph are restricted to have even degree. □

Assignment 11

Problem 1 (15pts)

Determine if the following statements are true or false. For each statement, briefly explain your reasoning.

(a) If Ham-Cycle is polynomial time reducible to interval scheduling problem then P = NP.

(b) The NP-Hard class of problems does not contain any decision problems.

(c) If there exists an algorithm that solves problem X with pseudo-polynomial runtime, then X must be NP-Hard.

(d) Suppose there is a 7-approxmiation algorithm for the general Traveling Salesman Problem. Then there exists a polynomial time solution for the 3-SAT problem.

(e) A vertex that is part of a minimum vertex cover can never be a part of a maximum independent set.

Solution.

(a) **True**. If $X \leq_p Y$, then we know we can solve X by solving Y and doing polynomial time additional work. We know from lecture that we can solve interval scheduling in $O(nlogn)$ time using a greedy algorithm. Therefore, if Ham-Cycle $\leq_p$ Interval Scheduling, we know we can solve Ham-Cycle by doing $O(nlogn)$ time to solve the Interval Scheduling problem and some polynomial p additional work. This results in polynomial total work to solve Ham-Cycle using Interval Scheduling. If Ham-Cycle $\in P$, then since Ham-Cycle is *NP-Complete*, we know the hardest problem in NP is $\in P$, so we know everything in NP is in P. Since every polynomial time algorithm has an efficient certificate, this completes the proof that Ham-Cycle $\leq_p$ Interval Scheduling $\implies P = NP$.

(b) **False.** By definition, all problems in *NP-Complete* are decision problems. Also by definition, all problems in *NP-Complete* are in NP-Hard. Therefore, we can take any problem in *NP-Complete*, such as 3-SAT from lecture, and it serves as a counterexample to the claim that NP-Hard doesn't contain any decision problems.

(c) **False.** Consider a problem which belongs to P such as the Max-Flow problem. We know that Edmonds-Karp finds Max-Flow in polynomial time. We also know that the unspecified Ford-Fulkerson finds Max-Flow in pseudo-polynomial runtime. Thus, Max-Flow is a problem X for which there exists an algorithm with pseudo-polynomial runtime. However, since we can solve Max-Flow in polynomial time, we know Max-Flow $\in P$, so we know Max-Flow $\notin$ NP-Hard. Thus, Max-Flow serves as a counterexample to the claim that the existence of a pseudo polynomial time algorithm for X implies X must be NP-Hard.

(d) **True.** Assuming the 7-approximation algorithm is itself polynomial, then we know the general TSP is polynomial. Since we know from lecture that the general TSP is in NP, this implies that $P = NP$. Since 3-SAT is in NP, this implies 3-SAT is in P. This proves that the existence of a polynomial time 7-approximation algorithm for the general TSP problem $\implies$ the existence of a polynomial time solution for the 3-SAT problem.

Note: if we do not assume that the approximation algorithm for the general TSP has polynomial runtime, then the answer is **false**. If the runtime of the 7-approximation algorithm is exponential, for example, then there might not be a polynomial time solution to 3-SAT, even though there is a 7-approximation algorithm for the general TSP.

(e) **False** Consider the graph $G = (V, E)$, where $V = \{A, B\}$ and $E = \{(a, b)\}$. Then there are two minimum vertex cover sets, $V_1 = \{A\}$ (since the only edge has one endpoint in $A($, and $V_2 = \{B\}$ (since the only edge has one endpoint in B). There are also two maximum independent sets, $V_3 = \{A\} = V_1$ and $V_4 = \{B\}$. Thus, $V_3 = V_1 = \{A\}$ is both a maximum independent set an a minimum vertex cover. Thus, the vertex A belongs to both a minimum vertex cover set and a maximum independent set. This disproves the claim that no vertex can belong to both a minimum vertex cover and a maximum independent set.

Problem 2 (15 pts)

Given an undirected graph with positive edge weights, the BIG-HAM-CYCLE problem is to decide if it contains a Hamiltonian cycle C such that the sum of weights of edges in C is at least half of the total sum of weights of edges in the graph. Show that finding BIG-HAM-CYCLE in a graph is NP-Complete.

Sodlution.
First, we must show that finding BIG-HAM-CYCLE is NP. In this case, the certificate is an ordered list of edges. There are $\leq |E|$ edges in this list, so the certificate is polynomial. The certifier must determine both

(i) if the certificate has exactly n edges in it, which it can do in constant time.

(ii) if the edges in the certificate pass through each node exactly once, which it can do in linear time.

Thus, the certifier takes linear time to verify the certificate, so it is polynomial with respect to the input length. Since BIG-HAM-CYCLE has both a polynomial length certificate and a polynomial time certifier, it has an efficient certification, so it is NP.

To show BIG-HAM-CYCLE is NP-Complete, we must reduce a known NP-Complete problem to a BIG-HAM-CYCLE problem in polynomial time. We will reduce the Ham-Cycle problem, which we know from lecture in NP-Complete. Consider any unweighted, undirected graph G. We need to construct a graph G' such that G' has a BIG-HAM-CYCLE $\iff$ G has a Ham-Cycle. By definition of a BIG-HAM-CYCLE C, the sum of the weights of edges in C must be *at least* half of the total sum of weights of edges in the graph. Therefore, if each edge e has weight w_e, we can write

$$\sum_{e \in C} w_e \geq \frac{\sum_{e \in E} w_e}{2}$$

Note: if $w_e = 0$ for all $e \in E$, than any subset of edges $S \subseteq E$ satisfies

$$\sum_{e \in S} w_e \geq \frac{\sum_{e \in E} w_e}{2}$$

Thus, we can construct G' as a weighted graph by setting $w_e = 0$ for all $e \in E$. Then, to determine if there is a Ham-Cycle in G, we can just use our blackbox to determine if there is a BIG-HAM-CYCLE in G'. It takes $O(|E|)$ time to set the edge weights of all edges to 0. Thus, we have reduced the known NP-Complete Ham-Cycle problem to the NP BIG-HAM-CYCLE problem in polynomial time. To show our reduction is valid, we must prove that there is a Ham-Cycle in $G \iff$ there is a BIG-HAM-CYCLE in G'.
If there is a Ham-Cycle C in G, then we can take the same set of edges C in G', and since we have not added any edges or nodes, we know C is a Ham-Cycle in G'. Also, since

$$\sum_{e \in C} w_e = 0 \geq \frac{0}{2} = \frac{\sum_{e \in E} w_e}{2}$$

we know C is a BIG-HAM-CYCLE in G'.
If we have a BIG-HAM-CYCLE C in G', then since we have not added or removed any nodes or edges, we know C is a Ham-Cycle in G.
This completes the proof that BIG-HAM-CYCLE is NP-Complete.

Problem 3 (15 pts)

Given an undirected connected graph $G = (V, E)$ in which a certain number of tokens $t(v) = 1$ or 2 placed on each vertex v. You will now play the following game. You pick a vertex u that contains at least two tokens, remove two tokens from u and add one token to any one of adjacent vertices. The objective of the game is to perform a sequence of moves such that you are left with exactly one token in the whole graph. You are not allowed to pick a vertex with 0 or 1 token. Prove that the problem of finding such a sequence of moves is NP-Hard by reduction from Hamiltonian Path.

Solution.
Consider a graph $G = (V, E)$. We need to construct a graph G' such that G has a Hamiltonian Path $\iff$ G' has a *winning sequence* of moves.
Note: If one vertex has two tokens, and all other vertices have one token, then a sequence of k moves (that passes through each vertex at most once) must leave k vertices with 0 tokens, 1 vertex with 2 tokens, and $n - k - 1$ vertices with 1 token (for all $1 \leq kleqn - 2$). *Proof.* We can induct on k.
Base Case: If $k = 1$, then we only have one move. The only legal first move is to move from the one vertex with 2 tokens to one of its neighbors. This causes the one vertex with 2 tokens to have 0 tokens, its neighbor to have 1+1 = 2 tokens, and the remaining n-1-1 = n-2 vertices to have 1 token.
Inductive Hypothesis: Assume the claim holds for all $1 \leq k \leq j$.
Inductive Step: Consider $k = j + 1$. By the inductive hypothesis, we know that there are j vertices with 0 tokens, 1 vertex with two tokens, and $n - j - 1$ vertices with 1 token. Therefore, the only legal $j + 1$th move is to go from the one vertex with two tokens to one of its neighbors with 1 token. This leaves $n - (j + 1) - 1 = n - j - 2$ vertices with 1 token, 1 vertex with 2 tokens, and $j + 1$ vertices with 0 tokens. The conclusion follows by induction.
We can use this result to construct G'. We just need to add a vertex $v*$ and edges connecting $v*$ to all vertices in V. That is, $G' = (V', E')$, where $V' = \{V \cup v*\}$ and $E' = E \cup \{(v*, v) | v \in V\}$. Then, following the intuition provided by the inductive result, we can set $t(v*) = 2$, and $t(v) = 1$ for all $v \in V$. To determine if there is a Hamiltonian Path in G, we can just use our blackbox to determine if there is a *winning sequence* in G'. It takes $O(1)$ time to add $v*$, $O(|E|)$ time to add all of the edges with endpoints in $v*$, and $O(|V|)$ time to set the token values for each vertex. Thus, we have reduced Hamiltonian Path to the token moving problem in polynomial time. To prove our reduction is valid, we need to show that there is a Hamiltonian Path in $G \iff$ there is a *winning strategy* to the token moving problem in G'.

If there is a Hamiltonian Path P in G, then we can go from $v*$ to either endpoint of P. This results in $v*$ having 0 tokens and the chosen endpoint of P having 2 tokens. By our inductive result, we know that tracing a sequence of n moves from $v*$ to one endpoint of P and then through P will result in one endpoint of P having 2 tokens and all other nodes having 0 tokens. To complete the winning strategy, we can simply move from the vertex with 2 tokens to $v*$. $v*$ will then be the only vertex with tokens, and it will only have 1 token.
If there is a *winning strategy* in G', we know it must start from $v*$ since that is the only legal first move. Since there is a winning strategy, we know all nodes from V were both moved to and from. Also we know that if we revisit a node $v \in V$ after moving to it once, we will leave that node with 1 token, so we will have no moves left. Thus, the only way for us to have a winning strategy is if we go from $v*$ to some $v \in V$ then through all other $u \in V$, then back to $v*$. This traces a Hamiltonian Cycle in G' which forms a Hamiltonian path in G when we exclude edges incident to $v*$ and the vertex $v*$ itself. Thus, the existence of a *winning strategy* in G' implies the existence of a Hamiltonian Path in G.
Thus, our polynomial time reduction from Hamiltonian Path to the token moving problem is valid.
This completes the proof that finding a *winning strategy* in the token moving problem is NP-Hard.

Problem 4 (20 pts)

In a certain town, there are many clubs, and every adult belongs to at least one club. The town's people would like to simplify their social life by disbanding as many clubs as possible, but they want to make sure that afterwards everyone will still belong to at least one club. Formally the Redundant Clubs problem has the following input and output. INPUT: List of people; list of clubs; list of members of each club; number K. OUTPUT: Yes if there exists a set of K clubs such that, after disbanding all clubs in this set, each person still belongs to at least one club. No otherwise. Prove that the Redundant Clubs problem is NP-Complete.

Solution.
First, we must show that the Redundant Clubs problem is NP. In this case, the certificate is a set of clubs to remove. If there are n clubs in the town, there are $\leq n$ clubs in the certificate, so it is polynomial length with respect to the input. The certifier must check both

(i) if the certificate is a set of size k, which can be done in $O(1)$ time.

(ii) if each person belongs to at least one club not in the set of size k, which can be done in polynomial time by brute force.

Thus, the Redundant Clubs problem has both a polynomial length certificate and a polynomial time certifier, so it has a n efficient certification. Thus, the Redundant Clubs problem is NP.
Now, we must reduce a known NP-Complete problem to the Redundant Clubs problem. We will reduce the Set Cover problem, which we know from lecture is NP-Complete. Consider an instance of Set Cover. We have a set P of p elements, a collection N of n subsets $N_i \subseteq P$, and we want to determine if there is a collection of k subsets whose union is all of P. This is essentially the complement of finding a collection of $n-k$ subsets $N_i \subseteq P$ such that removing those $n-k$ subsets still leaves each element of P in at least one of the remaining subsets. Thus, we can convert our Set Cover instance into a Redundant Clubs instance by letting P be our set of people and each N_i be the member list of one of the n clubs. To determine if there is a Set Cover of size k, we can simply use our blackbox to determine if there is a solution to the Redundant Clubs problem with the input $n-k$. It takes $O(p)$ time to copy the p elements into a list of people. It takes $O(np)$ time to copy the lists of club members for each club. Thus, we have reduced the Set Cover problem to the Redundant Clubs problem in polynomial time. To prove our reduction is valid, we must show that there is a Set Cover of size $k \iff$ there is a set of $n-k$ clubs that can be removed while maintaining that each person remains in at least one club.
If there is a set cover of size k, then we know the k corresponding clubs combine to have every single person in P as a member. Therefore, we know we can remove the other $n-k$ while guaranteeing that each person

remains in at least one club.
If there is a set of $n-k$ clubs that can be removed while maintaining that each person remains in at least one club, then we know the remaining k clubs must combine to have every single person in P as a member. Thus, the k corresponding subsets $N_i \subseteq P$ form a Set Cover of size k.
Thus, our polynomial time reduction from Set Cover to Redundant Clubs is valid.
This completes the proof that the Redundant Clubs is NP-Complete.

Problem 5 (15 pts)

Given a graph G=(V, E) with an even number of vertices as the input, the HALF-IS problem is to decide if G has an independent set of size $\frac{|V|}{2}$. Prove that HALF-IS is in NP-Complete.

Solution.
First, we must show that HALF-IS is NP. In this case, the certificate is a subset $S' \subseteq S$ of $\leq |V|$ vertices. Therefore, the certificate has polynomial length with respect to the input length. The certifier must both

(i) Verify that the set S' has size $\frac{|V|}{2}$, which can be done in constant time.

(ii) Verify that, for all $e \in E$, both endpoints of E are not in S', which can be done in $O(|E|log(|V|))$ time.

Thus, HALF-IS has both a polynomial length certificate and a polynomial time certifier, so it has an efficient certification. Thus, HALF-IS is NP.
Now, we must reduce a known NP-Complete problem to a HALF-IS problem. We will reduce the Independent Set problem, which we know from lecture is NP-Complete. Consider a graph $G = (V, E)$, on which we want to determine if there is an independent set of size k. Suppose $|V| = n$. Then if $k = \frac{n}{2}$, we can directly call our blackbox to determine if there is an independent set of size k. Otherwise, we need to construct $G' = (V', E')$ such that there is an independent set of size $\frac{|V'|}{2} \iff$ there is an independent set of size k in G. We do this in two different ways:

(i) If $k < \frac{n}{2}$, we need to add j nodes to V to form V' such that those j nodes can always be part of an independent set, and $\frac{k+j}{n+j} = \frac{1}{2}$. Solving this equation for j yields

$$j = n - 2k$$

Thus, if $k < \frac{n}{2}$, we construct G' with $V' = V \cup \{x_1, x_2, \cdots, x_{n-2k}\}$, and we let $E' = E$ so that $x_1, ..., x_{n-2k}$ can always be part of any independent set.

(ii) If $k > \frac{n}{2}$, we need to add j nodes to V to form V' such that those j nodes can never be part of an independent set of size > 1 and $\frac{k}{n+j} = \frac{1}{2}$. Solving this equation for j yields

$$j = 2k - n$$

Thus, if $k > \frac{n}{2}$, we construct G' with $V' = V \cup \{x_1, \cdots, x_{2k-n}\}$ and $E' = E \cup \{(x_i, x_j), (x_i, v) | 1 \leq i, j \leq 2k-n, v \in V\}$ so that $x_1, \cdots, x_{2k-n}$ can never be part of an independent set of size > 1.

Then, we can determine if there is an independent set of size k in G by using our blackbox to determine if there is an independent set of size $\frac{|V'|}{2}$ in G'. In both cases, we add $O(|V||E|)$ edges and $O(|V|)$ nodes to G to form G', so we have reduced the Independent Set problem to the HALF-IS problem in polynomial time. To show our reduction is valid, we need to show that G has an independent set of size $k \iff G'$ has an independent set of size $\frac{|V'|}{2}$.
If G has an independent set of size k, then

1. If $k < \frac{n}{2}$, we know that the union of the independent set of size k in G and the $x_1, \cdots, x_{n-2k}$ in G' form an independent set of size

$$n - k = \frac{n + n - 2k}{2} = \frac{2n + 2k}{2}$$

2. If $k > \frac{n}{2}$, we know that the independent set of size k forms an independent set of size

$$k = \frac{n + 2k - n}{2} = \frac{2k}{2}$$

So we can form an independent set of size $\frac{|V'|}{2}$ in both cases.
If we can form an independent set of size $\frac{|V'|}{2}$ in G', then

(i) If $k < \frac{n}{2}$, then if all of $x_1, \cdots, x_{n-2k}$ are in the independent set of size $\frac{|V'|}{2}$, we know removing them produces an independent set of size

$$n - k - (n - 2k) = k$$

in G. If not all of $x_1, \cdots, x_{n-2k}$ are in the independent set of size $\frac{|V'|}{2}$, then we know the set consists of more than k independent vertices from G, any subset of size k of which forms an independent set of size k in G.

(ii) If $k > \frac{n}{2}$, then we know the independent set of size $\frac{|V'|}{2}$ in G' consists only of nodes from G, and has size k, since none of the $x_1, \cdots, x_{2k-n}$ can be in an independent set of size > 1. Thus, the independent set of size $\frac{|V'|}{2}$ in G' itself forms an independent set of size k in G.

Thus, our polynomial time reduction from Independent Set to HALF-IS is valid.
This completes the proof that HALF-IS is NP-Complete.

Assignment 12

Problem 1 (25pts)

Suppose you are designing a scheduling algorithm for a manufacturing plant that has a set of M machines and a set of N tasks to be completed. Each task can be assigned to only one machine, and each machine can only perform one task at a time. Let p_{ij} be the processing time of task i on machine j. The goal is to minimize the overall time it takes to complete all tasks. How would you formulate this problem as an ILP? What decision variables would you use, and what constraints and the objective function would you include?

Solution.
For each task i, we need to decide whether or not we should complete it using machine j. Therefore, we can introduce

$$x_{ij} = \begin{cases} 1 \text{ if task } i \text{ is completed on task } j \\ 0 \text{ otherwise.} \end{cases}$$

Then, since we only need to complete each task once, we know $x_{ij} = 1$ for some $j = k$ and $x_{ij} = 0$ for all $\{j | 1 \leq j \neq k \leq M\}$. Thus, the total time to complete task i will be

$$\sum_{j=1}^{M} x_{ij} p_{ij}$$

Since different tasks can be completed consecutively on different machines, to minimize the total time it takes to complete all tasks, we must minimize the time at which the last task is completed. Let $C =$ the time at which the last task is completed. We know C will be equal to the amount of time taken by the machine which takes the most time. For each machine j, we know that the tasks assigned to that machine take a total of

$$\sum_{i=1}^{N} x_{ij} p_{ij}$$

time to complete. Thus, we know

$$C \geq \sum_{i=1}^{N} x_{ij} p_{ij}$$

for all $j \in \{1, ..., M\}$. In summary, we have x_{ij} as our decision variables, C as our objective function which we want to minimize, and our constraints are

$$\sum_{j=1}^{M} x_{ij} = 1 \text{ for all } i$$

$$x_{ij} \in \{0, 1\}$$

$$C \geq \sum_{i=1}^{N} x_{ij} p_{ij} \text{ for all } j$$

The first constraint ensures all tasks are assigned to exactly 1 machine. The second constraint ensures that the x_{ij}s function properly as decision variables corresponding to the assignment of task i to machine j. The third constraint ensures that no individual machine takes more time to complete its tasks than the total time to complete all tasks. These allow us to minimize the total time taken to complete all N tasks by using ILP to minimize C.

Problem 2 (25 pts)

Formulate the problem of finding a Min-S-T-cut of a directed network with source s and sink t as an Integer Linear Program and explain your program.

Solution.
For decision variables, we need to decide which nodes are in the S set and which nodes are in the T set. Thus, we can define

$$x_i = \begin{cases} 1 \text{ if } i \in S \\ 0 \text{ otherwise.} \end{cases}$$

for all $i \in V$.
We also need to determine which edges cross from the S cut to the T cut. Thus, we can define

$$y_{i,j} = \begin{cases} 1 \text{ if } (i,j) \text{ crosses the } S-T \text{ cut} \\ 0 \text{ otherwise} \end{cases}$$

for all $(i, j) \in E$.
We need to make sure that any edge (i, j) s.t. $i \in S$, $j \in T$ crosses the $S - T$ cut. Thus, we need

$$x_j - x_i + y_{i,j} \geq 0$$

for all $(i, j) \in E$. If $i, j \in S$ or $i, j \in T$, the statement is trivially true by the definitions of x_i, x_j, and $y_{i,j}$. If $j \in S$, $i \in T$, then the statement is trivially true. However, when $i \in S$, $j \in T$, we need $x_{i,j} = 1$ to satisfy the statement. This makes sense because any such (i, j) would have to cross the S-T cut. Since $x \in S$ and $t \in T$ by definition, we know $x_s = 1$ and $x_t = 0$.

To find the min S-T cut under these constraints, we minimize the objective function

$$C = \sum_{(i,j) \in E} c_{i,j} y_{i,j}$$

where $c_{i,j}$ is the capacity of the edge (i, j).

In summary, we can find the Min-S-T-cut of a directed network with source s and sink t by setting up the following integer linear programming problem:
Decision Variables: $y_{i,j}$ for all $(i,j) \in E$, x_i for all $i \in V$.
Objective Function: Minimize $C = \sum_{(i,j)\in E} c_{i,j}x_{i,j}$
Constraints:

(i) $x_s = 1$

(ii) $x_t = 0$

(iii) $x_i \in \{0,1\}$ for all $i \in (V - \{s,t\})$

(iv) $y_{i,j} \in \{0,1\}$ for all $(i,j) \in E$

(v) $x_j - x_i + y_{i,j} \geq 0$ for all $(i,j) \in E$

When minimizing C, we minimize the sum of the capacities of all edges crossing any S-T cut in G. Since the value of a cut is just the sum of the edges crossing it, minimizing C minimizes the value of any S-T cut in G. Thus, minimizing C finds the Min-S-T-Cut in G.

Problem 3 (25 pts)

A set of n space stations need your help in building a radar system to track spaceships traveling between them. The ith space station is located in 3D space at coordinates (x_i, y_i, z_i). The space stations never move. Each space station "i" will have a radar with power r_i, where r_i is to be determined. You want to figure out how powerful to make each space station's radar transmitter is, so that whenever any spaceship travels in a straight line from one station to another, it will always be in the radar range of either the first space station (its origin) or the second space station (its destination). A radar with power r is capable of tracking space ships anywhere in the sphere with radius r centered at itself. Thus, a spaceship is within radar range through its strip from space station i to space station j if every point along the line from (x_i, y_i, z_i) to (x_j, y_j, z_j) falls within either the sphere of radius r_i centered at (x_i, y_i, z_i) or the sphere of radius r_j centered at (x_j, y_j, z_j). The cost of each radar transmitter is proportional to its power, and you want to minimize the total cost of all of the radar transmitters. You are given all of the $(x_1, y_1, z_1), ..., (x_n, y_n, z_n)$ values, and your job is to choose values for $r_1, ..., r_n$. Express this problem as a linear program.

(a) Describe your variables for the linear program (5 pts).

(b) Write out the object function (8 pts).

(c) Describe the set of constraints for LP. You need to specify the number of constraints and describe what each constraint represents (12 pts).

Solution.

(a) The variables are the $r_1, ..., r_n$ for which we must determine values that minimize the total cost of all radar transmitters.

(b) We want to minimize the total cost of all radar transmitters. Since the cost of each radar transmitter i is proportional to its power r_i, we can minimize the total cost of all radar transmitters by minimizing

$$C = \sum_{i=1}^{n} r_i$$

(c) We know that any spaceship traveling in a straight line from station i to station j must be inside either the sphere of radius r_i centered at (x_i, y_i, z_i) or the sphere of radius r_j centered at (x_j, y_j, z_j) for all $i, j \in \{1, ..., n\}$. The distance between station i and station j is

$$D_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}$$

We know that any spaceship traveling from station i to station j can travel r_i distance along the line of distance D_{ij} while inside the sphere of radius r_i and r_j distance along the line of distance D_{ij} while inside the sphere of radius r_j. Thus, for the spaceship to be contained by either or both of the spheres during the entirety of its D distance journey from station i to station j, we need

$$r_i + r_j \geq D_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}$$

for all $i, j \in \{1, \cdots, n\}$. This is the constraint upon which we can minimize C.
Since there are $\binom{n}{2} = \frac{n(n-1)}{2}$ unique paths between the n stations, we have $\binom{n}{2}$ constraints, where each constraint corresponds to a unique path between stations.

Problem 4 (25 pts)

Recall the maximum-bipartite-matching problem. Write a linear program that solves this problem given a bipartite graph $G = (V, E)$, where the set of vertices on the left is L, and the set on the right is R, i.e. $L \cup R = V$.

Solution.
We will solve this problem using max-flow. Create a graph $G' = (V', E')$ such that $V' = s \cup V \cup t$ and $E' = \{(s, l_i) | l_i \in L\} \cup E \cup \{(r_i, t) | r_i \in R\}$ where all edges $e \in E'$ have capacity $c_e = 1$. Then the value of the max flow of G' equals the maximum size of a bipartite matching in G.

If G' has a max flow value of $v(f)$, then since all edges have capacity 1, we know we have $v(f)$ edge disjoint paths in G' such that no vertex (excluding s and t) has more than one path through it in the max flow f. Thus, we have $v(f)$ pairs (l_i, r_j) such that $l_i \in L$, $r_j \in R$, and the edge $(l_i, r_j) \in E$. Thus, we can create a matching of size $v(f)$ in G.
If G has a maximum bipartite matching size of $v(f)$, then we know there are $v(f)$ edges (l_i, r_j) s.t. $l_i \in L$, $r_i \in R$, and the l_is and r_is are all distinct. Since we have $(s, l_i) \in E'$ for all $l_i \in L$ and $(r_i, t) \in E'$ for all $r_i \in R$, we know we can create $v(f)$ edge disjoint paths in G' from s to t. Since each edge has capacity 1, we know each of these $v(f)$ paths will have a bottleneck of 1, so we know we can construct a flow of value $v(f)$ in G'.
This completes the proof that the value of max flow in G' is the maximum size of a bipartite matching in G.
Once we have made this reduction, we can easily construct a linear program to solve the problem. We have:

Variables: $f_{i,j}$ for all $(i, j) \in E'$
Objective Function: maximize flow $v(f) = \sum_{(s,l_i) \in E'} f_{s,l_i}$
Constraints: We need to add the typical constraints for a maximum flow problem. That is,

(i)
$$\sum_{(i,v) \in E'} f_{i,v} = \sum_{(v,j) \in E'} f_{v,j}$$

(ii) $0 \leq f_{i,j} \leq 1$ for all $(i, j) \in E'$

The first constraint maintains conservation of flow, while the second ensures that all edges e have non-negative flow that doesn't exceed c_e.
Since we know that maximizing $v(f)$ in G' maximizes the size of the bipartite matching in G, we can find the maximum size bipartite matching in G by using this linear program to maximize the s-t flow through G'

MATH 430: Theory of Numbers

All assignments in this section were written by Masoud Zargar, RTPC Assistant Professor of Mathematics, USC. Solutions to assignments 4 through 11 are provided.

Assignment 4

Problem 1

(6 Points). Using the Euclidean Algorithm, do the following:

(a) Find $gcd(53, 187)$ and find $x, y \in \mathbb{Z}$ such that $53x + 187y = gcd(53, 187)$.

Claim: $gcd(53, 187) = 1 = 53(60) + 187(-17)$, and $60, -17 = x, y \in \mathbb{Z}$ such that $53x + 187y = gcd(53, 187)$

Proof. We use the Euclidean Algorithm.

Note: $\forall a, b \in \mathbb{N}$, $gcd(a, b) = gcd(a, b - ka)$ $\forall k \in \mathbb{Z}$

Also Note: The division algorithm guarantees that $\forall a, b \in \mathbb{N}$, $a = bq + r$ for some unique $q, r \in \mathbb{Z}$, $0 \leq r < b$.

Now, we find:

187	$=3(53) + 28$	$\implies gcd(53, 187)$	$=gcd(53, 187 - 3(53))$	$=gcd(53, 28)$
53	$=1(28) + 25$	$\implies$	$=gcd(53 - 1(28), 28)$	$=gcd(25, 28)$
28	$=1(25) + 3$	$\implies$	$=gcd(25, 28 - 1(25)$	$=gcd(25, 3)$
25	$=8(3) + 1$	$\implies$	$=gcd(25 - 8(3), 3)$	$=gcd(1, 3)$
3	$=3(1) + 0$	$\implies$	$=gcd(1, 3 - 3(1))$	$=gcd(1, 0)$

$$
\begin{aligned}
\implies gcd(53, 187) = gcd(1, 0) = 1 &= 25 - 8(3) \\
&= 25 - 8(28 - 25) && = 9(25) - 8(28) \\
&= 9(53 - 28) - 8(28) && = 9(53) - 17(28) \\
&= 9(53) - 17(187 - 3(53)) && = 60(53) - 17(187)
\end{aligned}
$$

$\implies gcd(53, 187) = 1 = 53(60) + 187(-17)$
$\implies 60, -17 = x, y \in \mathbb{Z}$ such that $53x + 187y = 1 = gcd(53, 187)$ □

(b) Find $gcd(12345, 1234)$ and find $x, y \in \mathbb{Z}$ such that $12345x + 1234y = gcd(12345, 1234)$.

Claim: $gcd(12345, 1234) = 1 = 12345(247) + 1234(-2471)$, and $247, -2471 = x, y \in \mathbb{Z}$ such that 12345x + 1234y = gcd(12345,1234)

Proof. Similar to part (a), we use the Euclidean Algorithm.

We find:

$$\begin{array}{llllll} 12345 & =10(1234)+5 & \implies gcd(12345,1234) & =gcd(12345-10(1234),1234) & =gcd(5,1234) \\ 1234 & =246(5)+4 & \implies & =gcd(5,1234-246(5)) & =gcd(5,4) \\ 5 & =1(4)+1 & \implies & =gcd(5-1(4),4) & =gcd(1,4) \\ 4 & =4(1)+0 & \implies & =gcd(1,4-4(1)) & =gcd(1,0) \end{array}$$

$$\begin{array}{lllll} \implies gcd(12345,1234)=gcd(1,0)=1= & 5-4 & & & \\ & =5-(1234-246(5)) & = & 247(5)-1234 & \\ & =247(12345-10(1234))-1234 & = & 12345(247)+1234(-2471) & \end{array}$$

$\implies gcd(12345,1234)=1=12345(247)+1234(-2471)$
$\implies 247,-2471=x,y\in\mathbb{Z}$ such that $12345x+1234y=1=gcd(12345,1234)$ □

(c) Find $gcd(76,633)$ and find $x,y\in\mathbb{Z}$ such that $76x+633y=gcd(76,633)$.

Claim: $gcd(76,633)=1=76(25)+633(-3)$, and $25,-3=x,y\in\mathbb{Z}$ such that $76x+633y=gcd(76,633)$

Proof. Similar to parts (a) and (b), we use the Euclidean Algorithm.

We find:

$$\begin{array}{llllll} 633 & =8(76)+25 & \implies gcd(633,76) & =gcd(633-8(76),76) & =gcd(25,76) \\ 76 & =3(25)+1 & \implies & =gcd(25,76-3(25)) & =gcd(25,1) \\ 25 & =25(1)+0 & \implies & =gcd(25-25(1),1) & =gcd(0,1) \end{array}$$

$$\begin{array}{lllll} \implies gcd(76,633)=gcd(0,1)=1= & 76-3(25) & & & \\ & =76-3(633-8(76)) & = & 76(25)+633(-3) & \end{array}$$

$\implies gcd(76,633)=1=76(25)+633(-3)$
$\implies 25,-3=x,y\in\mathbb{Z}$ such that $76x+633y=gcd(76,633)$ □

Problem 2

(4 points). Show that if $gcd(a,b)=1$ for $a,b\in\mathbb{N}$, then

$$lcm(a^2+b^3,b^5+a^2b+b^4)=b(a^2+b^3)(b^4+a^2+b^3).$$

Proof. Note: $\forall a,b\in\mathbb{N}, lcm(a,b)=\frac{ab}{gcd(a,b)}$

Also note: $\forall a,b\in\mathbb{N},(a^2+b^3),(b^5+a^2b+b^4)\in\mathbb{N}$

$$\implies lcm(a^2+b^3,b^5+a^2b+b^4)=\frac{(a^2+b^3)(b^5+a^2b+b^4)}{gcd(a^2+b^3,b^5+a^2b+b^4)}=\frac{b(a^2+b^3)(b^4+a^2+b^3)}{gcd(a^2+b^3,b^5+a^2b+b^4)}$$

$\implies$ It is sufficient to show that $gcd(a^2+b^3, b^5+a^2b+b^4) = 1$

By Bézout's Theorem, since $gcd(a,b) = 1$, $\exists x, y \in \mathbb{Z}$ such that $1 = ax + by$. Squaring both sides, we get:

$$1^2 = 1 = (ax+by)^2 = a^2x^2 + 2axby + b^2y^2 = a^2(x^2) + b(2axy + by^2).$$

Since $a^2(x^2) + b(2axy + by^2) = 1$ is an integer linear combination of a^2 and b, $\implies gcd(a^2, b) = 1$.

Note: $\forall a, b \in \mathbb{N}$, $gcd(a,b) = gcd(a, b \pm ka)$, $\forall k \in \mathbb{Z}$

$\implies 1 = gcd(a^2, b) = gcd(a^2 + b^2(b), b) = gcd(a^2+b^3, b)$.

By Bézout's Theorem, since $gcd(a^2+b^3, b) = 1$, $\exists x_1, y_1 \in \mathbb{Z}$ such that $1 = (a^2+b^3)x_1 + by_1$.

Raising both sides to the fifth power, we get:

$$\begin{aligned} 1^5 = 1 = &\quad ((a^2+b^3)x_1 + by_1)^5 \\ = &\quad (a^2+b^3)^5x_1^5 + 5(a^2+b^3)^4x_1^4by_1 + 10(a^2+b^3)^3x_1^3b^2y_1^2 + 10(a^2+b^3)^2x_1^2b^3y_1^3 + 5(a^2+b^3)x_1b^4y_1^4 + b^5y_1^5 \\ = &\quad (a^2+b^3)((a^2+b^3)^4x_1^5 + 5(a^2+b^3)^3x_1^4by_1 + 10(a^2+b^3)^2x_1^3b^2y_1^2 + 10(a^2+b^3)x_1^2b^3y_1^3 + 5x_1b^4y_1^4) + b^5(y_1^5) \end{aligned}$$

Since $1 = (a^2+b^3)((a^2+b^3)^4x_1^5 + 5(a^2+b^3)^3x_1^4by_1 + 10(a^2+b^3)^2x_1^3b^2y_1^2 + 10(a^2+b^3)x_1^2b^3y_1^3 + 5x_1b^4y_1^4) + b^5(y_1^5)$

is an integer linear combination of (a^2+b^3) and b^5, $\implies gcd(a^2+b^3, b^5) = 1$

$\implies 1 = gcd(a^2+b^3, b^5) = gcd(a^2+b^3, b^5 + b(a^2+b^3)) = gcd(a^2+b^3, b^5+a^2b+b^4)$,

which is exactly what we want to show. Thus, $\forall a, b \in \mathbb{N}$ such that $gcd(a,b) = 1$

$$\begin{aligned} lcm(a^2+b^3, b^5+a^2b+b^4) = &\quad \frac{(a^2+b^3)(b^5+a^2b+b^4)}{gcd(a^2+b^3, b^5+a^2b+b^4)} &= &\quad \frac{(a^2+b^3)(b^5+a^2b+b^4)}{1} \\ = &\quad (a^2+b^3)(b^5+a^2b+b^4) &= &\quad b(a^2+b^3)(b^4+a^2+b^3) \end{aligned}$$

which concludes the proof.

□

Problem 3

(5 points). Suppose $a_1, ..., a_n$ are integers, at least one of which is nonzero. Prove that there are integers $x_1, ..., x_n$ such that

$$a_1x_1 + ... + a_nx_n = gcd(a_1, ..., a_n).$$

Proof. Let $S = \{a_1x_1 + ... + a_nx_n | x_1, ..., x_n \in \mathbb{Z}, a_1x_1 + ... + a_nx_n > 0\}$

Without loss of generality, assume $a_1 \neq 0$.

If $a_1 < 0$, then $x_1 = -1$, $x_2, ..., x_n = 0$ gives $a_1(-1) + 0(a_2 + ... + a_n) = |a_1| > 0 \in S$.

If $a_1 > 0$, then $x_1 = 1$, $x_2, ..., x_n = 0$ gives $a_1(1) + 0(a_2 + ... + a_n) = |a_1| > 0 \in S$.

$\implies S \neq \emptyset$.

Since all elements in S are greater than 0, the Well-Ordering Principle guarantees that S has a minimal element d.

We claim that $d = gcd(a_1, ..., a_n)$.

First, we must show $d|a_1, ..., a_n$, starting with a_1.

Note that the division algorithm guarantees $\exists$ unique $q, r \in \mathbb{Z}$ such that $a_1 = dq + r$, $0 \leq r < d$.

And $d \in S \implies d = a_1x_1 + ... + a_nx_n$, for some $x_1, ..., x_n \in \mathbb{Z}$, so $a_1 = (a_1x_1 + ... + a_nx_n)q + r$.

$\implies r = a_1(1 - x_1q) - a_2(qx_2) - ... - a_n(qx_n)$, so r is an integer linear combination of $a_1, ..., a_n$,

so $r > 0 \implies r \in S$, which contradicts the minimality of d.

Thus, $r = 0$, so $d|a_1$. We apply the exact same logic to show $d|a_2, ..., a_n$.

Thus, d is a common divisor of $a_1, ..., a_n$, and we need to show d is the *greatest* common divisor of $a_1, ..., a_n$.

Consider another common divisor of $a_1, ..., a_n$, which we'll call c.

Note: $c|a_1, ..., a_n \implies c|a_1x_1 + ... + a_nx_n = d \implies |c| \leq |d| = d$.

Therefore, $d = gcd(a_1, ..., a_n)$, and $d \in S$,

so we have proven that, if at least one of $a_1, ..., a_n \neq 0$, $\exists x_1, ..., x_n \in \mathbb{Z}$ such that

$$a_1x_1 + ... + a_nx_n = gcd(a_1, ..., a_n).$$

□

Problem 4

(5 points). Prove that $gcd(a, b, c) = gcd(gcd(a, b), c)$ from the basics.

Proof. Note: $gcd(a, b, c) \mid a, b \implies gcd(a, b, c) \mid ax + by = gcd(a, b)$ for some $x, y \in \mathbb{Z}$ (By Bézout's Theorem).

And we know $gcd(a, b, c) \mid c$, so we know $gcd(a, b, c)$ is a common divisor to $gcd(a, b)$ and c,

$\implies gcd(a, b, c) \leq gcd(gcd(a, b), c)$.

Similarly, $gcd(gcd(a, b), c) \mid gcd(a, b), c$, and we know $gcd(a, b) \mid a, b$, so we know $gcd(gcd(a, b), c) \mid a, b, c$.

$\implies gcd(gcd(a, b), c)$ is a common divisor to $a, b, c \implies gcd(gcd(a, b), c) \leq gcd(a, b, c)$.

Thus, $gcd(gcd(a, b), c) \leq gcd(a, b, c)$ and $gcd(a, b, c) \leq gcd(gcd(a, b), c)$, so we know $gcd(a, b, c) = gcd(gcd(a, b), c)$.

□

Problem 5

(Bonus, 5 points). Suppose s, t are distinct natural numbers such that

$$s^2 + st + t^2 \mid st(s + t)$$

Prove that $\mid s - t \mid \geq \sqrt[3]{st}$.

Proof. Note: $\forall k \in \mathbb{Z}$, $gcd(a,b) = gcd(a, b \pm ka)$. We'll call this property *Property 1.*

Also Note: By Bézout's Theorem, if $gcd(a,b) = 1$, $\exists x, y \in \mathbb{Z}$ such that $1 = ax + by$. Squaring both sides, we get:

$$1^2 = 1 = (ax + by)^2 = a^2x^2 + 2axby + b^2y^2 = a^2(x^2) + b(2axy + by^2).$$

Since $a^2(x^2) + b(2axy + by^2) = 1$ is an integer linear combination of a^2 and b, $\implies gcd(a^2, b) = 1$.

Therefore, $\forall a, b \in \mathbb{N}$, $gcd(a,b) = 1 \implies gcd(a^2, b) = 1$. We'll call this property *Property 2.*

Also Note: $(a \mid bc \ \&\ gcd(a,b) = 1) \implies a|c$. We'll call this property *Property 3.*

We will apply these 3 properties throughout the rest of the proof.

If we let $gcd(s,t) = d \in \mathbb{N}$, then we can write $s = dx$, $t = dy$ for some $x, y \in \mathbb{Z}$ such that $gcd(x,y) = 1$.

Substituting into the divisibility statement, we obtain:

$$\begin{aligned} (dx)^2 + (dx)(dy) + (dy)^2 = d^2(x^2 + xy + y^2) \mid \quad & (dx)(dy)(dx + dy) = d^3(xy)(x + y) \\ \implies x^2 + xy + y^2 \mid \quad & d(xy)(x + y) \end{aligned}$$

We know $gcd(x,y) = 1$, and *Property 1* tells us that $1 = gcd(x,y) = gcd(x, y + (1)x) = gcd(x, x + y)$.

Furthermore, *Property 2* tells us that $1 = gcd(x, x + y) = gcd(x^2, x + y)$.

Applying *Property 1* again, we find that $1 = gcd(x^2, x + y) = gcd(x^2 + y(x + y), x + y) = gcd(x^2 + xy + y^2, x + y)$.

Returning to our divisibility statement, we find that *Property 3* guarantees that

$x^2 + xy + y^2 \mid d(xy)$.

Similarly we can apply *Property 2* to find that

$1 = gcd(x^2 + xy + y^2, x + y) = gcd(x^2 + xy + y^2, (x + y)^2) = gcd(x^2 + xy + y^2, x^2 + 2xy + y^2)$.

Applying *Property 1* again, we find that

$1 = gcd(x^2 + xy + y^2, x^2 + 2xy + y^2) = gcd(x^2 + xy + y^2, x^2 + 2xy + y^2 - (x^2 + xy + y^2) = gcd(x^2 + xy + y^2, xy)$.

Returning once more to our divisibility statement, we find that *Property 3* guarantees that

$x^2 + xy + y^2 \mid d, \implies x^2 + xy + y^2 \leq d$

Now, we want to show that $|s - t| = |dx - dy| = d|x - y| \geq \sqrt[3]{st} = ((dx)(dy))^{1/3} = d^{2/3}(xy)^{1/3}$.

Note that this is true **if and only if** $d^{1/3}|x - y| \geq (xy)^{1/3}$

If $x > y$, we must show that $d^{1/3}(x - y) \geq (xy)^{1/3}$, which is true **if and only if** $d(x - y)^3 \geq xy$

Since $x^2 + xy + y^2 \leq d$, we know $d(x - y)^3 \geq (x^2 + xy + y^2)(x - y)^3$.

Since x, y are distinct natural numbers, and $x > y$, we know that $x - y \geq 1 \implies (x - y)^3 \geq 1$.

Therefore, $d(x-y)^3 \geq (x^2+xy+y^2)(x-y)^3 \geq (x^2+xy+y^2)$, and we know $x^2+xy+y^2 > xy$

since $x, y \in \mathbb{N} \implies x^2, y^2 > 0$.

Thus,

$$\begin{array}{rcccc} d(x-y)^3 & \geq & xy & & \\ \implies d^{1/3}(x-y) & \geq & (xy)^{1/3} & & \\ \implies d(x-y) & \geq & d^{2/3}(xy)^{1/3} & = & \sqrt[3]{d^2xy} \\ \implies (dx-dy) = s-t = |s-t| & \geq & \sqrt[3]{(dx)(dy)} & = & \sqrt[3]{st} \end{array}$$

which is exactly what we want to show.

Note: If $y > x$, we apply the exact same argument, but now $|x-y| = y-x$, so we just replace

every instance of $(x-y)$ with $(y-x)$, and the identical conclusion follows.

For clarity, we do that here:

If $y > x$, we must show that $d^{1/3}(y-x) \geq (xy)^{1/3}$, which is true **if and only if** $d(y-x)^3 \geq xy$

Since $x^2+xy+y^2 \leq d$, we know $d(y-x)^3 \geq (x^2+xy+y^2)(y-x)^3$.

Since x, y are distinct natural numbers, and $y > x$, we know that $y-x \geq 1 \implies (y-x)^3 \geq 1$.

Therefore, $d(y-x)^3 \geq (x^2+xy+y^2)(y-x)^3 \geq (x^2+xy+y^2)$, and we know $x^2+xy+y^2 > xy$

since $x, y \in \mathbb{N} \implies x^2, y^2 > 0$.

Thus,

$$\begin{array}{rcccc} d(y-x)^3 & \geq & xy & & \\ \implies d^{1/3}(y-x) & \geq & (xy)^{1/3} & & \\ \implies d(y-x) & \geq & d^{2/3}(xy)^{1/3} & = & \sqrt[3]{d^2xy} \\ \implies (dy-dx) = t-s = |s-t| & \geq & \sqrt[3]{(dx)(dy)} & = & \sqrt[3]{st} \end{array}$$

which is exactly what we want to show.

Thus, we have proven that, if s, t are distinct natural numbers such that $s^2+st+t^2 \mid st(s+t)$, then $|s-t| \geq \sqrt[3]{st}$.

□

Assignment 5

Problem 1

(6 points). (a) Find $gcd(56, 311)$ and find all solutions $(x, y) \in \mathbb{Z} \times \mathbb{Z}$ to the equation $56x + 311y = 3gcd(56, 311)$.

Claim: $gcd(56, 311) = 1$, and $x = 150 - 311t$, $y = -27 + 56t$, $t \in \mathbb{Z}$ represents all solutions

$(x, y) \in \mathbb{Z}$ x $\mathbb{Z}$ to the equation $56x + 311y = 3gcd(56, 311) = 3(1) = 3$

Proof. We use the Euclidean Algorithm.

Note: $\forall a, b \in \mathbb{N}$, $gcd(a, b) = gcd(a, b - ka)$ $\forall k \in \mathbb{Z}$

Now, we find:

311	$=5(56) + 31$	$\implies gcd(56, 311)$	$=gcd(56, 311 - 5(56))$	$=gcd(56, 31)$
56	$=1(31) + 25$	$\implies$	$=gcd(56 - 1(31), 31)$	$=gcd(25, 31)$
31	$=1(25) + 6$	$\implies$	$=gcd(25, 31 - 1(25))$	$=gcd(25, 6)$
25	$=4(6) + 1$	$\implies$	$=gcd(25 - 4(6), 6)$	$=gcd(1, 6)$
6	$=6(1) + 0$	$\implies$	$=gcd(1, 6 - 6(1))$	$=gcd(1, 0)$

$\implies gcd(56, 311) = gcd(1, 0) = 1.$

Note: Since $gcd(56, 311) = 1 | 3gcd(56, 311) = 3$, if we can find a single solution $(x_0, y_0) \in \mathbb{Z}$ x $\mathbb{Z}$ to the equation

$56x + 311y = 3gcd(56, 311) = 3(1) = 3$, then we know $x = x_0 - 311t$, $y = y_0 + 56t$, $t \in \mathbb{Z}$

represents all solutions $(x, y) \in \mathbb{Z}$ x $\mathbb{Z}$ to the equation $56x + 311y = 3gcd(56, 311) = 3(1) = 3$.

To find (x_0, y_0), note that:

$gcd(56, 311) = 1 =$	$25 - 4(6)$		
$=$	$25 - 4(31 - 25)$	$=$	$5(25) - 4(31)$
$=$	$5(56 - 31) - 4(31)$	$=$	$5(56) - 9(31)$
$=$	$5(56) - 9(311 - 5(56))$	$=$	$50(56) - 9(311)$

So, $56(50) + 311(-9) = gcd(56, 311) = 1 \implies 3(56(50) + 311(-9)) = 56(150) + 311(-27) = 3gcd(56, 311) = 3(1) = 3$

$\implies (150, -27) = (x_0, y_0) \in \mathbb{Z}$ x $\mathbb{Z}$

$\implies x = 150 - 311t$, $y = -27 + 56t$, $t \in \mathbb{Z}$ represents all solutions $(x, y) \in \mathbb{Z}$ x $\mathbb{Z}$ to the equation

$56x + 311y = 3gcd(56, 311) = 3$, which is exactly what we want to show, and thus concludes the proof.

□

(b) Find $gcd(42, 123)$ and find all solutions $(x, y) \in \mathbb{Z} \times \mathbb{Z}$ to the equation $42x + 123y = 5gcd(42, 123)$.

Claim: $gcd(42, 123) = 3$, and $x = 15 - 41t$, $y = -5 + 14t$, $t \in \mathbb{Z}$ represents all solutions

$(x, y) \in \mathbb{Z}$ x $\mathbb{Z}$ to the equation $42x + 123y = 5gcd(42, 123) = 5(3) = 15$

Proof. We use the Euclidean Algorithm. We find:

123	$=2(42) + 39$	$\implies gcd(42, 123)$	$=gcd(42, 123 - 2(42))$	$=gcd(42, 39)$
42	$=1(39) + 3$	$\implies$	$=gcd(42 - 1(39), 39)$	$=gcd(3, 39)$
39	$=13(3) + 0$	$\implies$	$=gcd(3, 39 - 13(3))$	$=gcd(3, 0)$

$\implies gcd(42, 123) = gcd(3, 0) = 3.$

Note: Since $gcd(42, 123) = 3|5gcd(42, 123) = 15$, if we can find a single solution $(x_0, y_0) \in \mathbb{Z}$ x $\mathbb{Z}$ to the equation $42x + 123y = 5gcd(42, 123) = 5(3) = 15$, then we know $x = x_0 - (123/3)t = x_0 - 41t$, $y = y_0 + (42/3)t = y_0 + 14t$, $t \in \mathbb{Z}$ represents all solutions $(x, y) \in \mathbb{Z}$ x $\mathbb{Z}$ to the equation $42x + 123y = 5gcd(42, 123) = 5(3) = 15$.

To find (x_0, y_0), note that:

$$\begin{aligned} gcd(42, 123) = 3 &= 42 - 39 \\ &= 42 - (123 - 2(42)) = 3(42) - 123 \end{aligned}$$

So, $42(3) + 123(-1) = gcd(42, 123) = 3 \implies 5(42(3) + 123(-1)) = 42(15) + 123(-5) = 5gcd(42, 123) = 5(3) = 15$

$\implies (15, -5) = (x_0, y_0) \in \mathbb{Z}$ x $\mathbb{Z}$

$\implies x = 15 - 41t$, $y = -5 + 14t$, $t \in \mathbb{Z}$ represents all solutions $(x, y) \in \mathbb{Z}$ x $\mathbb{Z}$ to the equation $42x + 123y = 5gcd(42, 123) = 15$, which is exactly what we want to show, and thus concludes the proof.

□

Problem 2

(4 points). Prove that if $n > 1$ is composite, then n has a prime factor not exceeding $\sqrt{n}$.

Proof. (By contradiction). Assume to the contrary that all of n's prime factors exceed $\sqrt{n}$, and let q be the number of n's prime factors (not necessarily distinct). Since n is composite, $q \geq 2$ (if $n = p^2$, p prime, $q = 2$ since $n = p * p$).

Note: $q \geq 2 \implies n = p_1...p_q$, where $p_1, ..., p_q > \sqrt{n}$ are (not necessarily distinct) primes $\implies n = p_1...p_q > \sqrt{n}\sqrt{n} = n$, which is a contradiction since $n = n$, so our initial assumption must be incorrect.

Thus, by contradiction, we have proven that if $n > 1$ is a composite, then n has a prime factor not exceeding $\sqrt{n}$.

□

Problem 3

(5 points). In this exercise, we show that:

$$S := 1 + \frac{1}{2} + ... + \frac{1}{n}$$

is never an integer when $n > 1$.

(a) Let $k \in \mathbb{N}$ be the largest power of 2 at most n so that $2^k \leq n < 2^{k+1}$. Let m be the least common multiple of $1, 2, 3, ..., n$, excluding 2^k. Show that $\frac{m}{2^k}$ is not an integer while $\frac{m}{1}, \frac{m}{2}, ..., \frac{m}{n}$, excluding $\frac{m}{2^k}$ are all integers.

Proof. Note: By the definition of the least common multiple, $1, 2, 3, ..., n$ (excluding 2^k) all divide m.

Therefore, $\frac{m}{1}, \frac{m}{2}, ..., \frac{m}{n}$, excluding $\frac{m}{2^k}$ are trivially all integers, so we just need to show that $\frac{m}{2^k} \notin \mathbb{Z}$.

Note that, since $2^k \leq n < 2^{k+1}$, 2^{k-1} is the largest power of 2 found in the unique prime factorizations of $1, 2, 3, ..., n$, excluding 2^k. To show this, consider the set $A := \{x \in \mathbb{N} | x \leq n, x \neq 2^k\}$. Since $2^k \notin A$, if 2^k was found in the unique prime factorization of any $a \in A$, then $a = p2^k$, where p is a prime. However, for all primes p, $p \geq 2$

$\implies a \geq 2(2^k) = 2^{k+1}$, which is a contradiction since, $\forall a \in A, a \leq n < 2^{k+1}$.

$\implies \forall a \in A$, 2^k is not in the unique prime factorization for a.

Thus, by unique prime factorization and the definition of least common multiple, $m = 2^{k-1} p_2^{a_2} ... p_j^{a_j}$, where $p_2, ..., p_j$ are distinct primes > 2 and $a_2, ..., a_j \in \mathbb{Z}, \geq 0$. Therefore, $\frac{m}{2^k} = \frac{p_2^{a_2} ... p_j^{a_j}}{2}$.

Since $p_2, ...p_j$ are distinct primes > 2, they are all odd, so $p_2^{a_2} ... p_j^{a_j}$ is odd, so $\frac{p_2^{a_2} ... p_j^{a_j}}{2} \notin \mathbb{Z}$.

Therefore, $\frac{m}{2^k} \notin \mathbb{Z}$, which is exactly what we want to show.

□

(b) Write:

$$mS = m + \frac{m}{2} + ... + \frac{m}{n},$$

And use the previous part to conclude that S is not an integer.

Proof. By the definition of the least common multiple, we know $m \in \mathbb{Z}$

From part (a), we know that $m, \frac{m}{2}, ..., \frac{m}{n}$ are all integers except $\frac{m}{2^k}$, which we know is not an integer.

Therefore, $b = m + \frac{m}{2} + ... + \frac{m}{n}$ (excluding $\frac{m}{2^k}$) is an integer.

Thus, $mS = b + \frac{m}{2^k}$. The sum of any integer with any non-integer is a non-integer, so we know $mS \notin \mathbb{Z}$.

Furthermore, the product of any two integers is an integer, and $m \in \mathbb{Z}$, so $S \notin \mathbb{Z}$ since $S \in \mathbb{Z} \implies mS \in \mathbb{Z}$.

Thus, $S := 1 + \frac{1}{2} + ... + \frac{1}{n} \notin \mathbb{Z}$, $\forall n \in \mathbb{N}, n > 1$, which concludes the proof.

□

Problem 4

(5 points). Suppose p and q are prime numbers, not necessarily distinct. Show that $\sqrt{p} + \sqrt{q} + \sqrt{pq}$ is irrational.

First, we must prove that, $\forall$ primes p, $\sqrt{p}$ is irrational.

Proof. (By Contradiction) Assume to the contrary that $\sqrt{p}$ is rational, that is, $\exists a, b \in \mathbb{N}$ such that $\sqrt{p} = \frac{a}{b}$.

$\implies b\sqrt{p} = a \implies b^2 p = a^2 \implies v_p(b^2 p) = v_p(a^2) \implies v_p(p) + v_p(b) + v_p(b) = v_p(a) + v_p(a)$.

$\implies 1 + 2v_p(b) = 2v_p(a)$, but $1 + 2v_p(b)$ is odd while $2v_p(a)$ is even, so we have a contradiction,

and our initial assumption must be incorrect. Thus, by contradiction, $\forall$ primes p, $\sqrt{p}$ is irrational.

□

We use this result to prove that $\sqrt{p} + \sqrt{q} + \sqrt{pq}$ is irrational.

Proof. (By Contradiction) Assume to the contrary that $\sqrt{p} + \sqrt{q} + \sqrt{pq}$ is rational, that is, $\exists c, d \in \mathbb{N}$ such that

$\sqrt{p} + \sqrt{q} + \sqrt{pq} = \frac{c}{d}$

$\implies \sqrt{p} + \sqrt{pq} = \frac{c}{d} - \sqrt{q} \implies p + 2p\sqrt{q} + pq = \frac{c^2}{d^2} - \frac{2c}{d}\sqrt{q} + q$

$\implies (2p + \frac{2c}{d})\sqrt{q} = \frac{c^2}{d^2} + q - p - pq$

$\implies \sqrt{q} = \frac{\frac{c^2}{d^2} + q - p - pq}{2p + \frac{2c}{d}} \implies \sqrt{q}$ is rational since $\frac{\frac{c^2}{d^2} + q - p - pq}{2p + \frac{2c}{d}}$ is rational (rational numbers are closed under addition,

multiplication, subtraction, and division), which is a contradiction since we know $\sqrt{q}$ is irrational since q is prime,

so our initial assumption must be incorrect.

Thus, by contradiction, $\sqrt{p} + \sqrt{q} + \sqrt{pq}$ is irrational $\forall$ primes p, q (not necessarily distinct).

□

Problem 5

(Bonus, 5 points). Show that

$$\sum_{n=0}^{\infty} \frac{1}{3^{n!}}$$

is irrational.

Proof. (By Contradiction). Assume to the contrary that $\sum_{n=0}^{\infty} \frac{1}{3^{n!}}$ is rational, that is, $\exists a, b \in \mathbb{N}$ such that

$$\sum_{n=0}^{\infty} \frac{1}{3^{n!}} = \frac{a}{b}$$

Note: $\forall q \in \mathbb{Q}$, $q = \frac{c}{d}$, for some $c, d \in \mathbb{N}$, such that $gcd(c, d) = 1$, $d \neq 0$.

Multiplying both sides by $1 = \frac{k}{k}$ yields $q\frac{k}{k} = q = \frac{kc}{kd}$, $\forall k \in \mathbb{Z}$.

Therefore, by letting $a = kc$, $b = kd$, we find that a and b can both be arbitrarily large natural numbers

if $\sum_{n=0}^{\infty} \frac{1}{3^{n!}} = \frac{a}{b} \in \mathbb{Q}$.

Also note: if $e, f \in \mathbb{N}$, then

$(e+f)! - e! = (e+f)(e+f-1)...(e+2)(e+1)e! - e! \geq (e+1)^f e! - e! \geq (e^f + 1)e! - e! = (e^f + 1 - 1)e! = e^f e!$

We will call this *Property 1*

Also Note: By definition, $\forall x \in \mathbb{N}$, $x! \geq 1 \implies x(x!) \geq x \implies 3^{x(x!)} \geq 3^x \implies \frac{x}{3^{x(x!)}-1} \leq \frac{x}{3^x-1}$.

Using L'Hopital's Rule, we find:

$$\lim_{x\to\infty} \frac{x}{3^{x(x!)}-1} \leq \lim_{x\to\infty} \frac{x}{3^x-1} = \lim_{x\to\infty} \frac{1}{ln(3)3^x} = 0 \implies 3^{x(x!)} - 1 > x \text{ for sufficiently large } x$$

since $\frac{x}{3^{x(x!)}-1} \geq 0 \ \forall x \in \mathbb{N} \implies \lim_{x\to\infty} \frac{x}{3^{x(x!)}-1} = 0$

We will call this property *Property 2*

Also Note: Using L'Hopital's Rule, we find, $\forall j \in \mathbb{N}$

$$\lim_{x\to\infty} \frac{jx}{x^j} = \lim_{x\to\infty} \frac{j}{jx^{j-1}} = \lim_{x\to\infty} \frac{1}{x^{j-1}} = 0 \implies x^j > jx \text{ for sufficiently large x}$$

We call this property *Property 3*

Now, consider:

$$S = b3^{b!}(\frac{a}{b} - \sum_{n=0}^{b} \frac{1}{3^{n!}}) = a3^{b!} - b\sum_{n=0}^{b} \frac{3^{b!}}{3^{n!}}$$

Since $a, 3^{b!} \in \mathbb{Z}$, $a3^{b!} \in \mathbb{Z}$

Also, since n ranges from 0 to b, $n! \leq b! \implies b! - n! \in \mathbb{Z}, \geq 0 \implies \frac{3^{b!}}{3^{n!}} = 3^{b!-n!} \in \mathbb{Z}$

$\implies b\sum_{n=0}^{b} \frac{3^{b!}}{3^{n!}} \in \mathbb{Z}$ since every term in the sum is an integer and b is an integer.

Therefore, S is the difference between two integers, so S is an integer.

Since $\forall x \in \mathbb{Z}$, $x \notin (0,1)$, showing that $0 < S < 1$ will yield a contradiction.

First, we show $S > 0$. Note that $S = b3^{b!}\sum_{n=b+1}^{\infty} \frac{1}{3^{n!}}$.

$b3^{b!} > 0$ since $b \in \mathbb{N}$, and $\frac{1}{3^{n!}} > 0 \ \forall n \geq b+1$, so every term in S is strictly greater than 0, so $S > 0$

Now, we show that $S < 1$. Note that:

$$S = b3^{b!}\sum_{n=b+1}^{\infty} \frac{1}{3^{n!}} = b\sum_{n=b+1}^{\infty} \frac{3^{b!}}{3^{n!}} = b\sum_{n=b+1}^{\infty} \frac{1}{3^{n!-b!}} = b(\frac{1}{3^{(b+1)!-b!}} + \frac{1}{3^{(b+2)!-b!}} + \frac{1}{3^{(b+3)!-b!}} + ...)$$

Applying *Property 1*, we find:

$$S \leq b(\frac{1}{3^{b!}} + \frac{1}{3^{b^2b!}} + \frac{1}{3^{b^3b!}} + ...) = b(\frac{1}{(3^{b!})^b} + \frac{1}{(3^{b!})^{b^2}} + \frac{1}{(3^{b!})^{b^3}} + ...)$$

Applying *Property 3*, and choosing b sufficiently large such that $b^j > jb \ \forall j \in \mathbb{N}$ we find:

$$S \leq b(\frac{1}{(3^{b!})^b} + \frac{1}{(3^{b!})^{2b}} + \frac{1}{(3^{b!})^{3b}} + ...) = b(\frac{1}{3^{b(b!)}} + \frac{1}{(3^{b(b!)})^2} + \frac{1}{(3^{b(b!)})^3} + ...) = \frac{b}{3^{b(b!)}}(1 + \frac{1}{3^{b(b!)}} + \frac{1}{(3^{b(b!)})^2} + ...)$$

Note: Since $(1 + \frac{1}{3^{b(b!)}} + \frac{1}{(3^{b(b!)})^2} + ...)$ is a geometric series,

$$(1 + \frac{1}{3^{b(b!)}} + \frac{1}{(3^{b(b!)})^2} + ...) = \frac{1}{1 - \frac{1}{3^{b(b!)}}} = \frac{1}{\frac{3^{b(b!)}}{3^{b(b!)}} - \frac{1}{3^{b(b!)}}} = \frac{1}{\frac{3^{b(b!)}-1}{3^{b(b!)}}} = \frac{3^{b(b!)}}{3^{b(b!)} - 1}$$

Therefore,

$$S \leq \frac{b}{3^{b(b!)}} \frac{3^{b(b!)}}{3^{b(b!)} - 1} = \frac{b}{3^{b(b!)} - 1}$$

From *Property 2*, we know $3^{b(b!)} - 1 > b$ for sufficiently large b.

Since b can be arbitrarily large, choose b sufficiently large such that $3^{b(b!)} - 1 > b$

Now, $\frac{b}{3^{b(b!)}-1} < 1$, and $S \leq \frac{b}{3^{b(b!)}-1}$, so we know $S < 1$, so we know $0 < S < 1$, which is a contradiction,

so our initial assumption must be incorrect.

Thus, by contradiction, we have proven that

$$\sum_{n=0}^{\infty} \frac{1}{3^{n!}}$$

is irrational. □

Assignment 6

Problem 1

(5 points). Suppose $a_1, ..., a_n \in \mathbb{Z}$ such that at least one of them is nonzero.

Show that $gcd(a_1, ..., a_n) = 1 \iff \exists\ x_1, ..., x_n \in \mathbb{Z}$ such that

$$a_1x_1 + ... + a_nx_n = 1$$

Proof. Note: Completing this proof requires both showing that

$(gcd(a_1, ..., a_n) = 1) \implies \exists\ x_1, ..., x_n \in \mathbb{Z}$ such that $a_1x_1 + ... + a_nx_n = 1$ and that

$(\exists\ x_1, ..., x_n \in \mathbb{Z}$ such that $a_1x_1 + ... + a_nx_n = 1) \implies gcd(a_1, ..., a_n) = 1$

From Problem 3 of Homework 4, we know that, since at least one of $a_1, ..., a_n$ is nonzero,

$\exists x_1, ..., x_n \in \mathbb{Z}$ such that $a_1x_1 + ... + a_nx_n = gcd(a_1, ..., a_n)$

Therefore, $(gcd(a_1, ..., a_n) = 1) \implies \exists\ x_1, ..., x_n \in \mathbb{Z}$ such that $a_1x_1 + ... + a_nx_n = gcd(a_1, ..., a_n) = 1$

Now, we just have to show that $(\exists\ x_1, ..., x_n \in \mathbb{Z}$ such that $a_1x_1 + ... + a_nx_n = 1) \implies gcd(a_1, ..., a_n) = 1$

By definition, $gcd(a_1, ..., a_n) \mid a_1, ..., a_n \implies gcd(a_1, ..., a_n) \mid a_1x_1 + ... + a_nx_n$

$\implies |gcd(a_1, ..., a_n)| = gcd(a_1, ..., a_n) \leq |a_1x_1 + ... + a_nx_n|$.

Since $gcd(a_1, .., a_n) \in \mathbb{N}$ by definition, we know that $gcd(a_1, ..., a_n) \geq 1$.

Therefore, $a_1x_1 + ... + a_nx_n = 1 \implies 1 \leq gcd(a_1, ..., a_n) \leq |1| = 1 \implies gcd(a_1, ..., a_n) = 1$.

Thus, $(gcd(a_1, ..., a_n) = 1) \implies \exists\ x_1, ..., x_n \in \mathbb{Z}$ such that $a_1x_1 + ... + a_nx_n = 1$ and

$(\exists\ x_1, ..., x_n \in \mathbb{Z}$ such that $a_1x_1 + ... + a_nx_n = 1) \implies gcd(a_1, ..., a_n) = 1$, so

$gcd(a_1, ..., a_n) = 1 \iff \exists\ x_1, ..., x_n \in \mathbb{Z}$ such that $a_1x_1 + ... + a_nx_n = 1$, which concludes the proof. □

Problem 2

(5 points). Compute $gcd(15, 21, 35)$ and find $x, y, z \in \mathbb{Z}$ such that

$$15x + 21y + 35z = 101gcd(15, 21, 35)$$

Claim: $gcd(15, 21, 35) = 1$, and $3636, -2424, -101$ is one solution $x, y, z \in Z$ such that $15x + 21y + 35z = 101gcd(15, 21, 35)$.

Proof. Note: $\forall a, b \in \mathbb{N}$, $gcd(a, b) = gcd(a, b - ka)\ \forall k \in \mathbb{Z}$

By Problem 4 of Homework 4, we know $gcd(15, 21, 35) = gcd(gcd(15, 21), 35)$.

To find $gcd(15, 21)$, we apply the Euclidean Algorithm and find:

$$\begin{aligned}
21 &= 1(15) + 6 &\implies\ & gcd(15, 21) &= gcd(15, 21 - 1(15)) &= gcd(15, 6)\\
15 &= 2(6) + 3 &\implies\ & &= gcd(15 - 2(6), 6) &= gcd(3, 6)\\
6 &= 2(3) + 0 &\implies\ & &= gcd(3, 6 - 2(3)) &= gcd(3, 0)
\end{aligned}$$

$\implies gcd(15, 21) = gcd(3, 0) = 3$

$\implies gcd(15, 21, 35) = gcd(3, 35) = gcd(3, 35 - 11(3)) = gcd(3, 2) = gcd(3 - 2, 2) = gcd(1, 2)$

$= gcd(1, 2 - 2(1)) = gcd(1, 0) = 1$

$\implies gcd(15, 21, 35) = gcd(1, 0) = 1 = 3 - 2 = 3 - (35 - 11(3)) = 12(3) - 35 = 12(15 - 2(6)) - 35$

$= 12(15) - 24(6) - 35 = 12(15) - 24(21 - 15) - 35 = 36(15) - 24(21) - 35$

$\implies 101gcd(15, 21, 35) = 101(1) = 101 = 101(36(15) - 24(21) - 35) = 3636(15) - 2424(21) - 101(35).$

$\implies 3636, -2424, -101 = x, y, z \in \mathbb{Z}$ such that $15x + 21y + 35z = 101gcd(15, 21, 35) = 101,$

which is exactly what we want to show, and thus concludes the proof. □

Problem 3

(5 points). Show that $\forall a, b, c \in \mathbb{N}$,

$$gcd(a, b, c)^2 \mid gcd(a, b)gcd(b, c)gcd(c, a)$$

Proof. Note: if $a_1 = p_1^{b_{1,1}}p_2^{b_{1,2}}...p_k^{b_{1,k}}, ..., a_n = p_1^{b_{n,1}}p_2^{b_{n,2}}...p_k^{b_{n,k}}$, then

$gcd(a_1, ..., a_n) = p_1^{min\{b_{1,1}, b_{2,1}, ..., b_{n,1}\}}...p_k^{min\{b_{1,k}, b_{2,k}, ..., b_{n,k}\}}$.

Also Note: Since $d|n \iff$ for all primes p, $v_p(d) \leq v_p(n)$, it suffices to show that, for all primes p,

$v_p(gcd(a, b, c)^2) \leq v_p(gcd(a, b)gcd(b, c)gcd(c, a))$

For the left hand side,

$v_p(gcd(a,b,c)^2) = v_p(gcd(a,b,c)) + v_p(gcd(a,b,c)) = 2v_p(gcd(a,b,c)) = 2min\{v_p(a), v_p(b), v_p(c)\}$

On the other hand, for the right hand side,

$v_p(gcd(a,b)gcd(b,c)gcd(c,a)) = v_p(gcd(a,b)) + v_p(gcd(b,c)) + v_p(gcd(c,a))$

$= min\{v_p(a), v_p(b)\} + min\{v_p(b), v_p(c)\} + min\{v_p(c), v_p(a)\}$. Since both of these sides are symmetric with respect to a, b, and c, we can assume, without loss of generality, that

$v_p(a) \leq v_p(b), v_p(c) \implies min\{v_p(a), v_p(b), v_p(c)\} = v_p(a)$.

Note that this implies that $v_p(a) = min\{v_p(a), v_p(b)\} = min\{v_p(c), v_p(a)\}$

Now, the left hand side of the inequality becomes $2v_p(a)$, while the right hand side becomes $2v_p(a) + min\{v_p(b), v_p(c)\}$

Since $min\{v_p(a), v_p(b), v_p(c)\} = v_p(a)$, we know $min\{v_p(b), v_p(c)\} \geq v_p(a) \implies 2v_p(a) + min\{v_p(b), v_p(c)\} \geq 3v_p(a)$

Combining the left and right hand sides, our inequality becomes

$v_p(gcd(a,b,c)^2) = 2v_p(a) \leq 3v_p(a) \leq 2v_p(a) + min\{v_p(b), v_p(c)\} = v_p(gcd(a,b)gcd(b,c)gcd(c,a))$

Since $a \in \mathbb{N}$, $v_p(a) \geq 0$, so $2v_p(a) \leq 3v_p(a)$ is true $\forall a$, so $v_p(gcd(a,b,c)^2) \leq v_p(gcd(a,b)gcd(b,c)gcd(c,a))$ $\forall a, b, c \in \mathbb{N}$,

and $\forall$ primes p, $\implies gcd(a,b,c)^2 \mid gcd(a,b)gcd(b,c)gcd(c,a)$ $\forall a, b, c \in \mathbb{N}$,

which is exactly what we want to show and thus concludes the proof. □

Problem 4

(5 points). For $n \in \mathbb{N}$, let

$$\psi(n) = \sum_{\substack{\alpha \in \mathbb{N} \\ p\ prime \\ p^\alpha \leq n}} log(p)$$

Show that

$$e^{\psi(n)} = lcm(1, 2, ..., n)$$

Proof. First, consider $a_1, ..., a_n \in \mathbb{N}$

We can write $a_1 = p_1^{c_{1,1}} ... p_k^{c_{1,k}}, ..., a_n = p_1^{c_{n,1}} ... p_k^{c_{n,k}}$ and $lcm(a_1, ..., a_n) = p_1^{d_1} ... p_k^{d_k}$,

where $p_1, ..., p_k$ are distinct primes and $c_{i,j}, d_j \geq 0, \in Z$, $\forall 1 \leq i \leq n$, $1 \leq j \leq k$.

By definition, $a_1, ..., a_n \mid lcm(a_1, ..., a_n) \implies d_j \geq max\{c_{1,j}, ..., c_{n,j}\}$, $\forall 1 \leq j \leq k$.

If, for any $1 \leq j \leq k$, $d_j > max\{c_{1,j}, ..., c_{n,j}\}$, then $\frac{lcm(a_1,...,a_n)}{p_j}$ would be a common multiple of $a_1, ..., a_n$ that is strictly less than $lcm(a_1, ..., a_n)$, which contradicts the definitional minimality of $lcm(a_1, ..., a_n)$.

Thus, $\forall 1 \leq j \leq k$, $d_j = max\{c_{1,j}, ..., c_{n,j}\}$.

Therefore, $lcm(a_1, ..., a_n) = p_1^{max\{c_{1,1},...,c_{n,1}\}}...p_k^{max\{c_{1,k},...,c_{n,k}\}}$.

If we let $(a_1, ..., a_n) = (1, 2, ..., n)$ and b_i = the maximum power of p_i that divides some element of $\{1, 2, ..., n\}$,

$\forall 1 \leq i \leq k$, then $lcm(a_1, ..., a_n) = lcm(1, 2, ..., n) = p_1^{b_1}p_2^{b_2}....p_k^{b_k}$

Now, we can analyze $\psi(n)$

$$\begin{aligned}\psi(n) = \sum_{\substack{\alpha\in\mathbb{N}\\ p\ prime\\ p^\alpha\leq n}} log(p) = & \underbrace{(log(p_1) + ... + log(p_1))}_{\alpha_1 \text{ times}} + \underbrace{(log(p_2) + ... + log(p_2))}_{\alpha_2 \text{ times}} + ... + \underbrace{(log(p_k) + ... + log(p_k))}_{\alpha_k \text{ times}} \\ = & \alpha_1 log(p_1) + \alpha_2 log(p_2) + ... + \alpha_k log(p_k) \\ = & log(p_1^{\alpha_1}) + log(p_2^{\alpha_2}) + ... + log(p_k^{\alpha_k}) \\ = & log(p_1^{\alpha_1}p_2^{\alpha_2}...p_k^{\alpha_k})\end{aligned}$$

$\implies e^{\psi(n)} = e^{log(p_1^{\alpha_1}p_2^{\alpha_2}...p_k^{\alpha_k})} = p_1^{\alpha_1}p_2^{\alpha_2}...p_k^{\alpha_k}$ where $p_1, ..., p_k$ are distinct primes as before.

Since α_i is the maximum power of p_i such that $p_i^{\alpha_i} \leq n$, α_i = the maximum power of p_i

that divides some element of $\{1, 2, ..., n\}$. (Since $p_i^{\alpha_i} \in \{1, 2, ..., n\}$ and, if $p_i^{\alpha_i+1}$ divides some element of $\{1, 2, ..., n\}$,

then $p_i^{\alpha_i+1} \leq n$, which contradicts the maximality of α_i.)

Therefore, $\forall 1 \leq i \leq k$, $\alpha_i = b_i$.

Therefore,

$$e^{\psi(n)} = p_1^{\alpha_1}p_2^{\alpha_2}...p_k^{\alpha_k} = p_1^{b_1}p_2^{b_2}...p_k^{b_k} = lcm(1, 2, ..., n)$$

which is exactly what we want to show, and thus concludes the proof. □

Problem 5

(Bonus, 5 points). Suppose s, t are distinct natural numbers. Show that

$$lcm(s, t) + lcm(s + 1, t + 1) > \frac{2st}{\sqrt{|s - t|}}$$

Proof. First, since the inequality is symmetric with respect to s and t, we can assume, without loss of generality,

that $s > t$, meaning $\sqrt{|s - t|} = \sqrt{s - t}$.

Note: $\forall n \in \mathbb{N}$, $gcd(n, n + 1) = gcd(n, n + 1 - n) = gcd(n, 1) = gcd(n - n(1), 1) = gcd(0, 1) = 1$,

so all consecutive natural numbers are coprime. Since $gcd(s, t) = gcd(s - t, t)$, we know $gcd(s, t)|s - t, t$.

Similarly, $gcd(s + 1, t + 1) = gcd(s + 1 - (t + 1), t + 1) = gcd(s - t, t + 1) \implies gcd(s + 1, t + 1)|s - t, t + 1$

If $d := gcd(gcd(s, t), gcd(s + 1, t + 1)) > 1$, then, for some $x, y \in \mathbb{N}$, $t = xd, t + 1 = yd \implies gcd(t, t + 1) \geq d > 1$,

which is a contradiction since $\forall t \in \mathbb{N}$, $gcd(t, t + 1) = 1$. Therefore, $d := gcd(gcd(s, t), gcd(s + 1, t + 1)) = 1$

Since $gcd(s, t), gcd(s + 1, t + 1)|s - t$, and $gcd(gcd(s, t), gcd(s + 1, t + 1)) = 1$, we know $gcd(s, t)gcd(s + 1, t + 1)|s - t$

$\implies |gcd(s,t)gcd(s+1,t+1)| = gcd(s,t)gcd(s+1,t+1) \leq s-t = |s-t|$. We will call this *Property 1.*

Also note that, $\forall a,b \in \mathbb{N}$, $lcm(a,b) = \frac{ab}{gcd(a,b)} \implies lcm(s,t) = \frac{st}{gcd(s,t)}$, and $lcm(s+1,t+1) = \frac{(s+1)(t+1)}{gcd(s+1,t+1)}$.

Therefore, it is sufficient to show that:

$$\begin{aligned} & \frac{st}{gcd(s,t)} + \frac{(s+1)(t+1)}{gcd(s+1,t+1)} && > && \frac{2st}{\sqrt{|s-t|}} \\ \iff \quad & \frac{stgcd(s+1,t+1) + (s+1)(t+1)gcd(s,t)}{gcd(s,t)gcd(s+1,t+1)} && > && \frac{2st}{\sqrt{s-t}} \qquad (\sqrt{|s-t|} = \sqrt{s-t} \text{ since } s > t.) \end{aligned}$$

Note: $(s+1)(t+1) > st$

$$\implies \frac{stgcd(s+1,t+1) + (s+1)(t+1)gcd(s,t)}{gcd(s,t)gcd(s+1,t+1)} > \frac{stgcd(s+1,t+1) + stgcd(s,t)}{gcd(s,t)gcd(s+1,t+1)}$$

So it is sufficient to show that:

$$\frac{stgcd(s+1,t+1) + stgcd(s,t)}{gcd(s,t)gcd(s+1,t+1)} = \frac{st(gcd(s+1,t+1) + gcd(s,t))}{gcd(s,t)gcd(s+1,t+1)} \geq \frac{2st}{\sqrt{s-t}}$$

Dividing both sides of the inequality by st yields a new inequality we need to prove:

$$\frac{gcd(s+1,t+1) + gcd(s,t)}{gcd(s,t)gcd(s+1,t+1)} \geq \frac{2}{\sqrt{s-t}}$$

Note: By the AM-GM inequality, we know that $\forall a,b \in \mathbb{N}$, $\frac{a+b}{2} \geq \sqrt{ab} \implies a+b \geq 2\sqrt{ab}$.

By definition of the greatest common divisor, we know $gcd(s,t), gcd(s+1,t+1) \in \mathbb{N}$

$\implies gcd(s,t) + gcd(s+1,t+1) \geq 2\sqrt{gcd(s,t)gcd(s+1,t+1)}$. Therefore, we know:

$$\frac{gcd(s+1,t+1) + gcd(s,t)}{gcd(s,t)gcd(s+1,t+1)} \geq \frac{2\sqrt{gcd(s,t)gcd(s+1,t+1)}}{gcd(s,t)gcd(s+1,t+1)} = \frac{2}{\sqrt{gcd(s,t)gcd(s+1,t+1)}}$$

And we want to show that $\frac{2}{\sqrt{gcd(s,t)gcd(s+1,t+1)}} \geq \frac{2}{\sqrt{s-t}}$. Note:

$$\begin{aligned} & \frac{2}{\sqrt{gcd(s,t)gcd(s+1,t+1)}} && \geq && \frac{2}{\sqrt{s-t}} \\ \iff \quad & 2\sqrt{s-t} && \geq && 2\sqrt{gcd(s,t)gcd(s+1,t+1)} \\ \iff \quad & \sqrt{s-t} && \geq && \sqrt{gcd(s,t)gcd(s+1,t+1)} \\ \iff \quad & s-t && \geq && gcd(s,t)gcd(s+1,t+1) \end{aligned}$$

By *Property 1*, we already know that $s-t \geq gcd(s,t)gcd(s+1,t+1)$. Therefore, we also know:

$$\begin{aligned} & \frac{2}{\sqrt{gcd(s,t)gcd(s+1,t+1)}} && \geq && \frac{2}{\sqrt{s-t}} \\ \implies \quad & \frac{gcd(s+1,t+1) + gcd(s,t)}{gcd(s,t)gcd(s+1,t+1)} && \geq && \frac{2\sqrt{gcd(s,t)gcd(s+1,t+1)}}{gcd(s,t)gcd(s+1,t+1)} = \frac{2}{\sqrt{gcd(s,t)gcd(s+1,t+1)}} \geq \frac{2}{\sqrt{s-t}} \end{aligned}$$

Multiplying the leftmost and rightmost terms by st yields:

$$\implies st\frac{gcd(s+1,t+1) + gcd(s,t)}{gcd(s,t)gcd(s+1,t+1)} = \frac{stgcd(s+1,t+1) + stgcd(s,t)}{gcd(s,t)gcd(s+1,t+1)} \geq st\frac{2}{\sqrt{s-t}} = \frac{2st}{\sqrt{s-t}}$$

Since $st < (s+1)(t+1)$, we also know:

$$\frac{stgcd(s+1,t+1)+(s+1)(t+1)gcd(s,t)}{gcd(s,t)gcd(s+1,t+1)} > \frac{stgcd(s+1,t+1)+stgcd(s,t)}{gcd(s,t)gcd(s+1,t+1)} \geq \frac{2st}{\sqrt{s-t}}$$

Note:

$$\frac{stgcd(s+1,t+1)+(s+1)(t+1)gcd(s,t)}{gcd(s,t)gcd(s+1,t+1)} = \frac{st}{gcd(s,t)} + \frac{(s+1)(t+1)}{gcd(s+1,t+1)} = lcm(s,t)+lcm(s+1,t+1)$$

Thus, we have shown that:

$$lcm(s,t)+lcm(s+1,t+1) > \frac{stgcd(s+1,t+1)+stgcd(s,t)}{gcd(s,t)gcd(s+1,t+1)} \geq \frac{2st}{\sqrt{s-t}} = \frac{2st}{\sqrt{|s-t|}}$$

which is exactly what we want to show, and thus concludes the proof.

□

Assignment 7

Problem 1

(5 points). Solve the following system of congruences: $\begin{cases} x \equiv 20 \pmod{19} \\ x \equiv 3 \pmod{23} \\ x \equiv 4 \pmod{5} \end{cases}$

Note that since 19, 23, and 5 are all prime, 19, 23, and 5 are all coprime, and so the Chinese Remainder Theorem guarantees the existence of a unique solution $x \pmod{(19)(23)(5)}$ to the system.

To find x, we will apply the Chinese Remainder Theorem:

First, let $a_1 = 20, a_2 = 3, a_3 = 4, n_1 = 19, n_2 = 23, n_3 = 5, N_1 = n_2n_3 = 23(5), N_2 = n_1n_3 = 19(5), N_3 = n_1n_2 = 19(23)$

We want to find x_1 such that $N_1x_1 \equiv 1 \pmod{19}$.

Note: $N_1 = 23(5) \equiv 4(5) = 20 \equiv 1 \pmod{19} \implies N_1x_1 \equiv x_1 \equiv 1 \pmod{19}$, so we can take $x_1 = 1$.

Now, we want to find x_2 such that $N_2x_2 \equiv 1 \pmod{23}$.

Note: $N_2 = 19(5) = 95 \equiv 3 \pmod{23} \implies N_2x_2 \equiv 3x_2 \equiv 1 \pmod{23} \implies 8(3x_2) = 24x_2 \equiv x_2 \equiv 8(1) = 8 \pmod{23}$,

so we can take $x_2 = 8$.

Now, we want to find x_3 such that $N_3x_3 \equiv 1 \pmod{5}$.

Note: $N_3 = 19(23) \equiv (4)(3) = 12 \equiv 2 \pmod{5} \implies N_3x_3 \equiv 2x_3 \equiv 1 \pmod{5}$

$\implies 3(2x_3) = 6x_3 \equiv x_3 \equiv 3(1) = 3 \pmod{5}$, so we can take $x_3 = 3$.

Now, the Chinese Remainder Theorem guarantees that

$$\begin{aligned} x = &\ a_1N_1x_1 + a_2N_2x_2 + a_3N_3x_3 = 20(23)(5)(1) + 3(19)(5)(8) + 4(19)(23)(3) = 9824 = 4(19)(23)(5) + 1084 \\ \equiv &\ 1084 \pmod{(19)(23)(5)} \end{aligned}$$

is the unique solution $\pmod{(19)(23)(5)}$ to the system of congruences.

Problem 2

(5 points). Solve the following system of congruences: (*System 1*) $\begin{cases} 2x \equiv 11 \pmod{17} \\ 3x \equiv 12 \pmod{21} \\ 4x \equiv 13 \pmod{31} \end{cases}$

Note: $\begin{cases} 2x \equiv 11 \pmod{17} \iff 9(2x) = 18x \equiv x \equiv 9(11) = 99 \equiv 14 \pmod{17} \\ 3x \equiv 12 \pmod{21} \iff 3x - 12 = 21y \iff x - 4 = 7y \iff x \equiv 4 \pmod{7} \\ 4x \equiv 13 \pmod{31} \iff 8(4x) = 32x \equiv x \equiv 8(13) = 104 \equiv 11 \pmod{31} \end{cases}$

Now we can define another system: (*System 2*) $\begin{cases} x \equiv 14 \pmod{17} \\ x \equiv 4 \pmod{7} \\ x \equiv 11 \pmod{31} \end{cases}$

Note: Since 17, 7, and 31 are all prime, they are all coprime, so the Chinese Remainder Theorem

guarantees the existence of a unique solution $x \pmod{(17)(7)(31)}$ to *System 2*.

System 1 and *System 2* are logically equivalent, and the Chinese Remainder Theorem guarantees the existence of a unique solution $x \pmod{(17)(7)(31)}$ to *System 2*, so it also guarantees the existence of that same unique solution $x \pmod{(17)(7)(31)}$ to *System 1*. To find x, we will apply the Chinese Remainder Theorem.

First, let $a_1 = 14, a_2 = 4, a_3 = 11, n_1 = 17, n_2 = 7, n_3 = 31,$

$N_1 = n_2 n_3 = 7(31), N_2 = n_1 n_3 = 17(31), N_3 = n_1 n_2 = 17(7).$

We want to find x_1 such that $N_1 x_1 \equiv 1 \pmod{17}$. Note: $N_1 = 7(31) \equiv 7(14) = 98 \equiv 13 \pmod{17}$

$\implies N_1 x_1 \equiv 13x_1 \equiv 1 \pmod{17} \implies 4(13x_1) = 52x_1 \equiv x_1 \equiv 4(1) = 4 \pmod{17}$, so we can take $x_1 = 4$.

Now, we want to find x_2 such that $N_2 x_2 \equiv 1 \pmod{7}$. Note: $N_2 = 17(31) \equiv 3(3) = 9 \equiv 2 \pmod{7}$

$\implies N_2 x_2 \equiv 2x_2 \equiv 1 \pmod{7} \implies 4(2x_2) = 8x_2 \equiv x_2 \equiv 4(1) = 4 \pmod{7}$, so we can take $x_2 = 4$.

Now, we want to find x_3 such that $N_3 x_3 \equiv 1 \pmod{31}$. Note: $N_3 = 17(7) = 119 \equiv 26 \pmod{31}$

$\implies N_3 x_3 \equiv 26x_3 \equiv 1 \pmod{31} \implies 6(26x_3) = 156x_3 \equiv x_3 \equiv 6(1) = 6 \pmod{31}$, so we can take $x_3 = 6$.

Now, the Chinese Remainder Theorem guarantees that

$$\begin{aligned} x = &\ a_1 N_1 x_1 + a_2 N_2 x_2 + a_3 N_3 x_3 = 14(7)(31)(4) + 4(17)(31)(4) + 11(17)(7)(6) = 28438 = 7(17)(7)(31) + 2615 \\ \equiv &\ 2615 \pmod{(17)(7)(31)} \end{aligned}$$

is the unique solution $\pmod{(17)(7)(31)}$ to both *System 2* and *System 1*.

Problem 3

(5 points). A man had a near fatal encounter, as a result of which he lost most of his memory. An FBI agent pays him a visit and asks him what he was doing the previous day. The man says that he only remembers that he walked a certain distance (denote it by d meters). He remembers that, for some reason, when d was divided into 1 kilometer pieces, 53 meters remained. When he divided it into 13 meters, 3 meters remained.

He knows also that he walked less than 15 kilometers. How many meters d did the man walk? Perhaps a mathematician is involved...
Note that we can solve for d by solving the following system of congruences, under the condition $0 < d < 15000$:

$\begin{cases} d \equiv 53 \pmod{1000} \\ d \equiv 3 \pmod{13} \end{cases}$ Note: 13 is prime, and $1000 = 10(10)(10) = (5)(2)(5)(2)(5)(2) = 2^3 5^3$,

so 13 and 1000 are coprime. Therefore, the Chinese Remainder Theorem guarantees the existence of a unqiue solution

$x \pmod{(13)(1000)}$ to the system. To find x, we apply the Chinese Remainder Theorem.

First, let $a_1 = 53, a_2 = 3, n_1 = 1000, n_2 = 13, N_1 = n_2 = 13, N_2 = n_1 = 1000$

We want to find x_1 such that $N_1 x_1 \equiv 1 \pmod{1000}$. Note: $N_1 = 13 \equiv 13 \pmod{1000}$

$\implies N_1 x_1 = 13x_1 \equiv 1 \pmod{1000} \implies 77(13x_1) = 1001x_1 \equiv x_1 \equiv 77(1) = 77 \pmod{1000}$, so we can take $x_1 = 77$.

Now, we want to find x_2 such that $N_2 x_2 \equiv 1 \pmod{13}$. Note: $N_2 = 1000 \equiv 12 \pmod{13}$

$\implies N_2 x_2 \equiv 12x_2 \equiv -x_2 \equiv 1 \pmod{13} \implies x_2 \equiv -1 \equiv 12 \pmod{13}$, so we can take $x_2 = 12$

Now, the Chinese Remainder Theorem guarantees that

$$\begin{aligned} x = & \quad a_1 N_1 x_1 + a_2 N_2 x_2 = 53(13)(77) + 3(1000)(12) = 89053 = 6(13)(1000) + 11053 \\ \equiv & \quad 11053 \pmod{(13)(1000)} \end{aligned}$$

is the unique solution $\pmod{(13)(1000)}$ to the system of congruences.

Since $11053 \equiv 11053 \pmod{(13)(1000)}$, and since $0 < 11053 < 15000$, $d = 11053$,

so the man must have walked 11053 meters.

Problem 4

(5 points). Find $2^{2022} \pmod{21}$

Claim: $2^{2022} \equiv 1 \pmod{21}$

Proof. Let $r \equiv 2^{2022} \pmod{21}$. Since $21 = 7(3)$, this yields the system of congruences:

(System 1) $\begin{cases} r \equiv 2^{2022} \pmod{7} \\ r \equiv 2^{2022} \pmod{3} \end{cases}$

Since 7 and 3 are coprime, the Chinese Remainder Theorem guarantees the existence of a unique solution

$r \pmod{(7)(3)=21}$ to *System 1.*

By Fermat's Little Theorem, for all natural numbers n and primes p such that $gcd(n, p) = 1$, $n^{p-1} \equiv 1 \pmod{p}$.

$gcd(2, 7) = 1 \implies 2^6 \equiv 1 \pmod{7} \implies r \equiv 2^{2022} = (2^6)^{337} \equiv 1^{337} = 1 \pmod{7}$

Similarly, $gcd(2, 3) = 1 \implies 2^2 \equiv 1 \pmod{3} \implies r \equiv 2^{2022} = (2^2)^{1011} \equiv 1^{1011} = 1 \pmod{3}$

Since $1 \equiv 1 \pmod 7$, and $1 \equiv 1 \pmod 3$, $r = 1 \equiv 1 \pmod{21}$ is one solution to *System 1.*

Since the Chinese Remainder Theorem guarantees the existence of a *unique* solution (mod 21) to *System 1*,

we know $r = 1 \equiv 1 (\text{mod } 21)$ is the unique solution (mod 21) to *System 1.*

$\implies r \equiv 2^{2022} \equiv 1 \pmod{21}$, which is exactly what we want to show, and thus concludes the proof. $\square$

Problem 5

(Bonus, 5 points). Show that for every prime p, there is an $n \in \mathbb{N}$ such that

$$\frac{2^n + 3^n + 6^n - 1}{p^3} \in \mathbb{N}$$

Proof. First, note that, since $p > 0$, $\forall$ primes p, and since $2^n + 3^n + 6^n - 1 > 0$, $\forall n \in \mathbb{N}$, we know that

$$\frac{2^n + 3^n + 6^n - 1}{p^3} \in \mathbb{N} \iff 2^n + 3^n + 6^n - 1 \equiv 0 \pmod{p^3}$$

Also note that, by Euler's Theorem, $\forall a \in \mathbb{N}$, p prime, if $gcd(a, p^3) = 1$, then $a^{\varphi(p^3)} \equiv 1 \pmod{p^3}$,

where $\varphi(p^3) = p^2(p-1)$.

If $p = 2$, then we can take $n = 2$, and

$$2^2 + 3^2 + 6^2 - 1 = 48 = 6(2^3) \equiv 0 \pmod{2^3} \iff \frac{2^2 + 3^2 + 6^2 - 1}{2^3} = \frac{48}{8} = 6 \in \mathbb{N}$$

so we know $n = 2$ is a suitable $n \in \mathbb{N}$ for $p = 2$.

If $p = 3$, we want to find a suitable n such that

$$2^n + 3^n + 6^n - 1 \equiv 0 \pmod{3^3}$$

Note that $6^n = 2^n 3^n$, and let $n = 3x$ for some $x \in \mathbb{N}$, so we can write

$2^n + 3^n + 6^n - 1 = 2^{3x} + (3^3)^x + 2^{3x}(3^3)^x - 1 \equiv 2^{3x} + 0^x + 2^{3x}0^x - 1 \equiv 2^{3x} - 1 \equiv 0 \pmod{3^3} \iff 2^n = 2^{3x} \equiv 1 \pmod{3^3}$

By Euler's Theorem, we know $2^{\varphi(3^3)} = 2^{3^2(2)} = 2^{18} \equiv 1 \pmod{3^3}$, and $\varphi(3^3) = 18$ is a positive multiple of 3,

allowing $x = 6 \in \mathbb{N}$, so we know $n = \varphi(3^3) = 18$ is a suitable $n \in \mathbb{N}$ for $p = 3$

To be sure, we can plug 18 in for n, and we find

$$\frac{2^{18} + 3^{18} + 6^{18} - 1}{3^3} = \frac{262144 + 387420489 + 10155995666846 - 1}{27} = \frac{101560344351048}{27} = 3761494235224 \in \mathbb{N}$$

as expected, so we know $n = 18$ is a suitable $n \in \mathbb{N}$ for $p = 3$.

Now, we just need to consider primes $p > 3$.

Note that, since $6^n = 2^n 3^n$, $\forall$ primes $p > 3$, $n \in \mathbb{N}$, $gcd(2^n, p^3) = gcd(3^n, p^3) = gcd(6^n, p^3) = 1$.

Therefore, by Euler's Theorem, for all primes $p > 3$,

$$2^{\varphi(p^3)} \equiv 3^{\varphi(p^3)} \equiv 6^{\varphi(p^3)} \equiv 1 \pmod{p^3}$$

Let $n = \varphi(p^3) - 1$, and we find

$$2^n+3^n+6^n-1 = 2^{\varphi(p^3)-1}+3^{\varphi(p^3)-1}+6^{\varphi(p^3)-1}-1 = \frac{2^{\varphi(p^3)}}{2}+\frac{3^{\varphi(p^3)}}{3}+\frac{6^{\varphi(p^3)}}{6}-1 \equiv \frac{1}{2}+\frac{1}{3}+\frac{1}{6}-1 = 1-1 = 0 \pmod{p^3}$$

$$\iff \frac{2^{\varphi(p^3)-1} + 3^{\varphi(p^3)-1} + 6^{\varphi(p^3)-1} - 1}{p^3} \in \mathbb{N}$$

so we know $n = \varphi(p^3) - 1$ is a suitable $n \in \mathbb{N}$ for all primes $p > 3$.

Since we already found suitable values of $n \in \mathbb{N}$ for all primes $p \leq 3$ (just 2 and 3), we have found a suitable value of $n \in \mathbb{N}$ such that

$$\frac{2^n + 3^n + 6^n - 1}{p^3} \in \mathbb{N}$$

for all primes p, which is exactly what we want to show, and thus concludes the proof.

□

Assignment 8

Problem 1

(4 points). Show that if $p > 1$ is a natural number such that

$$(p-1)! \equiv -1 \pmod{p}$$

then p is a prime number.

Proof. (By Contradiction) Assume to the contrary that $p > 1$ is a natural number such that

$$(p-1)! \equiv -1 \pmod{p}$$

and p is a composite number.

Since p is composite, we know $\exists k \in \{2, ..., p-1\}$ such that $p = ks$ for some $s \in \mathbb{N}$.

Since $(p-1)! \equiv -1 \pmod{p}$, we know $(p-1)! + 1 = pr = ksr$ for some $r \in \mathbb{Z}$.

$\implies (p-1)! + 1$ is a multiple of k $\implies (p-1)! + 1 \equiv 0 \pmod{k} \implies (p-1)! \equiv -1 \pmod{k}$.

However, since $k \in \{2, ..., p-1\}$, $(p-1)! = 1...(k-1)(k)(k+1)...(p-1)$,

so $(p-1)!$ is a multiple of k, so $(p-1)! \equiv 0 \pmod{k}$, which is a contradiction.

Thus, our initial assumption that p is composite must be false, so, if $p > 1$ is a natural number such that

$$(p-1)! \equiv -1 \pmod{p}$$

then p must be a prime number, which is exactly what we want to show, and thus concludes the proof. □

Problem 2

(6 points). Find the last two digits of 3^{1000}. Find the last two digits of 2^{1000}.

First, note that, by Euler's Theorem, $\forall a \in \mathbb{N}$, p prime, if $gcd(a, p^n) = 1$, then $a^{\varphi(p^n)} \equiv 1 \pmod{p^n}$,

where $\varphi(p^n) = p^{n-1}(p-1)$. We will use this frequently throughout the rest of the problem.

Claim: The last two digits of 3^{1000} are 01.

Proof. Note that finding the last two digits of 3^{1000} is the same as solving for $r \equiv 3^{1000} \pmod{100}$.

Since $100 = 5^2 2^2$, this yields the following system of congruences: *(System 1)* $\begin{cases} r \equiv 3^{1000} \pmod{5^2} \\ r \equiv 3^{1000} \pmod{2^2} \end{cases}$ Since $gcd(5^2, 2^2) = 1$, the Chinese

to *System 1*.

By Euler's Theorem, since $gcd(3, 5^2) = 1$, we know that $3^{\varphi(5^2)} = 3^{5^1(5-1)} = 3^{20} \equiv 1 \pmod{5^2}$

$\implies r \equiv 3^{1000} = (3^{20})^{50} \equiv 1^{50} = 1 \pmod{5^2}$.

Also by Euler's Theorem, since $gcd(3, 2^2) = 1$, we know that $3^{\varphi(2^2)} = 3^{2^1(2-1)} = 3^2 \equiv 1 \pmod{2^2}$.

$\implies r \equiv 3^{1000} = (3^2)^{500} \equiv 1^{500} = 1 \pmod{2^2}$.

Since $1 \equiv 1 \pmod{5^2}$ and $1 \equiv 1 \pmod{2^2}$, we know $r = 1 \equiv 1 \pmod{(5^2)(2^2) = 100}$ is a solution to *System 1*.

Since the Chinese Remainder Theorem guarantees the existence of a *unique* solution (mod $(5^2)(2^2)$) to *System 1*,

we know $r = 1 \equiv 1 \pmod{100}$ is the unique solution (mod 100) to *System 1*.

$\implies r \equiv 3^{1000} \equiv 1 \pmod{100}$, so the last two digits of 3^{1000} are 01, which is exactly what we want to show,

and thus concludes the proof. □

Claim: The last two digits of 2^{1000} are 76.

Proof. Note that finding the last two digits of 2^{1000} is the same as solving for $s \equiv 2^{1000} \pmod{100}$.

Since $100 = 5^2 2^2$, this yields the following system of congruences: *(System 2)* $\begin{cases} s \equiv 2^{1000} \pmod{5^2} \\ s \equiv 2^{1000} \pmod{2^2} \end{cases}$ Since $gcd(2^2, 5^2) = 1$, the Chinese

to *System 2*.

By Euler's Theorem, since $gcd(2, 5^2) = 1$, we know that $2^{\varphi(5^2)} = 2^{5^1(5-1)} = 2^{20} \equiv 1 \pmod{5^2}$.

$\implies s \equiv 2^{1000} = (2^{20})^{50} \equiv 1^{50} = 1 \pmod{5^2}$.

Also note: $s \equiv 2^{1000} = (2^2)^{500} \equiv 0^{500} = 0 \pmod{2^2}$,

so we need to find an s such that $s \equiv 1 \pmod{5^2}$ and $s \equiv 0 \pmod{2^2}$.

We only care about the last two digits of s, so we only consider the potential values of s (mod 100),

which are 0 through 99 (inclusive). Of these, only 1, 26, 51, and $76 \equiv 1 \pmod{5^2}$. Of these, only $76 \equiv 0 \pmod{2^2}$.

$\implies s = 76 \equiv 76 \pmod{(5^2)(2^2)}$ is a solution (mod 100) to *System 2*.

Since the Chinese Remainder Theorem guarantees the existence of a *unique* solution (mod $(5^2)(2^2) = 100$) to *System 2*,

we know $s = 76 \equiv 76 \pmod{100}$ is the unique solution (mod 100) to *System 2*.

$\implies s \equiv 2^{1000} \equiv 76 \pmod{100}$, so the last two digits of 2^{1000} are 76, which is exactly what we want to show,

and thus concludes the proof.

□

Problem 3

(5 points). Show that for any pair of natural numbers m, n such that $gcd(m, n) = 1$,

$$\varphi(mn) = \varphi(m)\varphi(n)$$

Proof. Note: if $m = 1$ or $n = 1$ or $m = n = 1$, $gcd(m, n)$ always equals 1, and

$$\varphi(mn) = \varphi(m)\varphi(n)$$

is trivially true since $\varphi(1) = 1$, so we just need to show the equation holds for $m, n \neq 1 \implies m, n \geq 2$.

Also note: $\forall\ a \in \mathbb{N}$, $a \geq 2$, $a = p_1^{\alpha_1}...p_k^{\alpha_k}$, where $p_1, ..., p_k$ are *distinct* primes, $\alpha_1, ..., \alpha_k \in \mathbb{Z}, \geq 1$, and $k \in \mathbb{N}$.

For all such $a \in \mathbb{N}$, $a \geq 2$, $\varphi(a) = a(1 - \frac{1}{p_1})...(1 - \frac{1}{p_k})$

Using these facts, write $m = p_{m,1}^{\alpha_1}...p_{m,k}^{\alpha_k}$ and $n = p_{n,1}^{\beta_1}...p_{n,t}^{\beta_t}$, where $p_{m,1}, ..., p_{m,k}$ are distinct primes,

$p_{n,1}, ..., p_{n,t}$ are distinct primes, $\alpha_i \in \mathbb{Z}, \geq 1$ for all $1 \leq i \leq k$, and $\beta_j \in \mathbb{Z}, \geq 1$ for all $1 \leq j \leq t$.

By the definition of the Euler Totient,

$$\varphi(m) = m(1 - \frac{1}{p_{m,1}})...(1 - \frac{1}{p_{m,k}})$$

and

$$\varphi(n) = n(1 - \frac{1}{p_{n,1}})...(1 - \frac{1}{p_{n,t}})$$

$$\implies \varphi(m)\varphi(n) = m(1-\frac{1}{p_{m,1}})...(1-\frac{1}{p_{m,k}})n(1-\frac{1}{p_{n,1}})...(1-\frac{1}{p_{n,t}}) = mn(1-\frac{1}{p_{m,1}})...(1-\frac{1}{p_{m,k}})(1-\frac{1}{p_{n,1}})...(1-\frac{1}{p_{n,t}})$$

Since $gcd(m, n) = 1$, m and n share no prime factors, so $p_{m,i} \neq p_{n,j}$ for all $1 \leq i \leq k$, $1 \leq j \leq t$.

Since $m = p_{m,1}^{\alpha_1}...p_{m,k}^{\alpha_k}$ and $n = p_{n,1}^{\beta_1}...p_{n,t}^{\beta_t}$, we know

$$mn = p_{m,1}^{\alpha_1}...p_{m,k}^{\alpha_k}p_{n,1}^{\beta_1}...p_{n,t}^{\beta_t}$$

where $p_{m,1}, ..., p_{m,k}, p_{n,1}, ..., p_{n,t}$ are all distinct prime numbers.

Therefore, by the above formula for the Euler Totient,

$$\varphi(mn) = mn(1-\frac{1}{p_{m,1}})...(1-\frac{1}{p_{m,k}})(1-\frac{1}{p_{n,1}})...(1-\frac{1}{p_{n,t}}) = m(1-\frac{1}{p_{m,1}})...(1-\frac{1}{p_{m,k}})n(1-\frac{1}{p_{n,1}})...(1-\frac{1}{p_{n,t}}) = \varphi(m)\varphi(n)$$

for all $m, n \in \mathbb{N}, \geq 2$ such that $gcd(m, n) = 1$.

Since $\varphi(mn) = \varphi(m)\varphi(n)$ is trivially true if $m = 1$, or $n = 1$, or $m = n = 1$ such that $gcd(m, n) = 1$,

and we have shown that $\varphi(mn) = \varphi(m)\varphi(n)$ for all $m, n \in \mathbb{N}, \geq 2$ such that $gcd(m, n) = 1$, we know that

$$\varphi(mn) = \varphi(m)\varphi(n)$$

for all $m, n \in \mathbb{N}$ such that $gcd(m, n) = 1$, which is exactly what we want to show, and thus concludes the proof. □

Problem 4

(5 points). Use orders of elements to show that for any pair of natural numbers $a > n$,

$$n|\varphi(a^n - 1)$$

Proof. Note: $gcd(a, b) = gcd(a, b - ka)\ \forall\ a, b, k \in \mathbb{Z}$

Therefore, $gcd(a, a^n - 1) = gcd(a, a^n - 1 - a^{n-1}a) = gcd(a, -1) = gcd(a - (-a)(-1), -1) = gcd(0, -1) = 1$

By Euler's Theorem, we know that, $\forall\ a, b \in \mathbb{N}$ such that $gcd(a, b) = 1, a^{\varphi(b)} \equiv 1 \pmod{b}$.

Since $a, n \in \mathbb{N}$, and $a > n \implies a \geq 2$, we know that $a, a^n - 1 \in \mathbb{N}$.

Combining this with Euler's Theorem and the fact that $gcd(a, a^n - 1) = 1$, we know that $a^{\varphi(a^n-1)} \equiv 1 \pmod{a^n - 1}$

Since $a^n = (a^n - 1) + 1$, we know that $a^n \equiv 1 \pmod{a^n - 1}$ as well.

Since $a^m \equiv 1 \pmod{a^n - 1} \implies ord_{a^n-1}(a)|m$, it suffices to show that $n = ord_{a^n-1}(a)$.

To do so, we assume to the contrary that $n \neq ord_{a^n-1}(a)$, that is, $\exists k \in \mathbb{N}$ such that $a^k \equiv 1 \pmod{a^n - 1}$ and $k < n$.

Since $a \geq 2$ and $k \in \mathbb{N} \geq 1$, $a^k \geq 2$, so $a^k \equiv 1 \pmod{a^n - 1} \implies a^k - 1 = (a^n - 1)t$ for some $t \in \mathbb{N}$.

However, $k < n \implies a^k - 1 < a^n - 1 \implies \frac{a^k-1}{a^n-1} = t < 1$, which is a contradiction since $t \in \mathbb{N} \geq 1$.

Thus, our initial assumption must be incorrect, so we know $n = ord_{a^n-1}(a)$,

$\implies n|\varphi(a^n - 1)$ for all pairs of natural numbers $a > n$, which concludes the proof. □

Problem 5

(Bonus, 5 points). Show that there is no natural number $n > 1$ such that

$$n|2^n - 1$$

Proof. (By Contradiction). Assume to the contrary that $\exists$ a natural number $n > 1$ such that $n|2^n - 1$.

If n is even, $n|2^n - 1 \implies 2^n - 1 = na$ for some $a \in \mathbb{Z}$. But 2^n is even $\implies 2^n - 1$ is odd,

and n is even $\implies na$ is even $\forall a \in \mathbb{Z}$, so we have a contradiction.

If n is odd, since $n > 1$, n is either an odd prime or the product of odd primes.

If n is an odd prime, Fermat's Little Theorem guarantees that $2^n \equiv 2 \pmod{n}$,

but $n|2^n - 1 \implies 2^n - 1 = nk$ for some $k \in \mathbb{Z} \implies 2^n - 1 \equiv 0 \pmod{n} \implies 2^n \equiv 1 \pmod{n}$, and $1 \not\equiv 2 \pmod{n}$,

for all $n > 1$, so we have a contradiction.

Finally, if n is a product of odd primes, then we can let p be the smallest such prime, that is, $n = px$ for some $x \in \mathbb{N}$.

Since p is n's smallest prime factor, $gcd(n, p-1) = 1$ (since the prime factorization of $p - 1$ only contains

primes that are strictly less than p).

Since p is an odd prime, $gcd(2, p) = 1$, so Fermat's Little Theorem guarantees that $2^{p-1} \equiv 1 \pmod{p}$.

Since $n|2^n - 1$, $2^n - 1 = ns = pxs$ for some $s \in \mathbb{Z} \implies 2^n - 1 \equiv 0 \pmod{p} \implies 2^n \equiv 1 \equiv 2^{p-1} \pmod{p}$.

Let $ord_p(2) = t \in \mathbb{N}$ be the smallest power of 2 such that $2^t \equiv 1 \pmod{p}$.

Since $2^1 = 2 \not\equiv 1 \pmod{p}$ $\forall$ odd primes p, we know that $t > 1$.

Since $2^n \equiv 2^{p-1} \equiv 1 \pmod{p}$, we know that $t|n$ and $t|p-1$

$\implies t|gcd(n, p-1) \implies |t| = t \leq |gcd(n, p-1)| = gcd(n, p-1) = 1.$

But $t > 1$, so we have a contradiction.

Thus, for all natural numbers $n > 1$, assuming that $n|2^n - 1$ yields a contradiction.

Therefore, our assumption must be incorrect, so we have shown that there is no natural number $n > 1$ such that

$$n|2^n - 1$$

which is exactly what we want to show, and thus concludes the proof. □

Assignment 9

Problem 1

(5 points). (a) Suppose G is a group. Show that if e_1, e_2 are both units in G, then $e_1 = e_2$, that is, units are unique.

Proof. By the definition of a group, if e is a unit in a group G, then, for all $g \in G$, $e * g = g * e = g$.

Therefore, since e_1 is a unit in G, $e_1 * e_2 = e_2$.

Similarly, since e_2 is a unit in G, $e_1 * e_2 = e_1$.

Therefore, $e_1 = e_1 * e_2 = e_2 \implies e_1 = e_2$, which is exactly what we want to show,

and thus concludes the proof. □

(b) Suppose $x \in G$, and that y, $z \in G$ are inverses of x. Show that $y = z$, that is, inverses are unique.

Proof. By the definition of a group, we know that, for all $g \in G$, if $a \in G$ is an inverse to g,

then $a * g = g * a = e$, where e is the (unique by part (a)) unit for G.

Therefore, $x * y = y * x = e = z * x = x * z$.

For all $g \in G$, $e * g = g * e = g$, so $y = y * e$. But $e = x * z$.

$\implies y = y * e = y * (x * z) = (y * x) * z = e * z = z$, which is exactly what we want to show,

and thus concludes the proof. □

Problem 2

(5 points). Let

$$\tau(n) := \sum_{d|n, d \in \mathbb{N}} 1$$

be the number of positive divisors of n. Systematically compute $\tau(175)$, $\tau(231)$.

Let $p_1, ..., p_k$ be distinct primes. Find $\tau(p_1 p_2^2 p_3^3 ... p_k^k)$ in terms of k.

Note: $\forall n \in \mathbb{N}, \geq 2$, we can write $n = p_1^{\alpha_1} ... p_k^{\alpha_k}$, where $p_1, ..., p_k$ are distinct primes, and $\alpha_i \in \mathbb{N}$ for all $1 \leq i \leq k$.

For all such n, $\tau(n) = (\alpha_1 + 1)...(\alpha_k + 1)$. We will call this *Property 1*, and refer back to it

throughout the rest of this problem.

Claim: $\tau(175) = 6$

Proof. $175 \geq 2$, and we can write $175 = p_1^{\alpha_1} p_2^{\alpha_2} = 5^2 7^1 \implies \alpha_1 = 2$, $\alpha_2 = 1$.

By *Property 1*, $\implies \tau(175) = (\alpha_1 + 1)(\alpha_2 + 1) = (2 + 1)(1 + 1) = 3(2) = 6$, which is exactly what we want to show,

and thus concludes the proof. □

Claim: $\tau(231) = 8$.

Proof. $231 \geq 2$, and we can write $231 = p_1^{\alpha_1} p_2^{\alpha_2} p_3^{\alpha_3} = 3^1 7^1 11^1 \implies \alpha_1 = \alpha_2 = \alpha_3 = 1$

By *Property 1*, $\implies \tau(231) = (\alpha_1 + 1)(\alpha_2 + 1)(\alpha_3 + 1) = (1 + 1)(1 + 1)(1 + 1) = 2^3 = 8$, which is exactly what

we want to show, and thus concludes the proof. □

Claim: $\tau(p_1 p_2^2 p_3^3 ... p_k^k) = (k + 1)!$.

Proof. Since $n = p_1 p_2^2 p_3^3 ... p_k^k$ is the product of at least one prime, and all primes are ≥ 2, we know that $n \geq 2$.

Moreover, $n = p_1 p_2^2 p_3^3 ... p_k^k = p_1^{\alpha_1} p_2^{\alpha_2} p_3^{\alpha_3} ... p_k^{\alpha_k} \implies \alpha_i = i$, for all $1 \leq i \leq k$.

By *Property 1*,

$$\begin{aligned} \implies \tau(n) = \tau(p_1 p_2^2 p_3^3 ... p_k^k) = & \quad (\alpha_1 + 1)(\alpha_2 + 1)(\alpha_3 + 1)...(\alpha_k + 1) = (1 + 1)(2 + 1)(3 + 1)...(k + 1) \\ = & \quad 2(3)(4)...(k + 1) = 1(2)(3)(4)...(k + 1) = (k + 1)!, \end{aligned}$$

which is exactly what we want to show, and thus concludes the proof. □

Problem 3

(5 points). Let

$$\sigma(n) := \sum_{d|n, d\in\mathbb{N}} d$$

be the sum of divisors of n. Compute $\sigma(175)$, $\sigma(231)$. Find the sum of the *even* positive divisors of 10000.

Note: $\forall n \in \mathbb{N}, \geq 2$, we can write $n = p_1^{\alpha_1}...p_k^{\alpha_k}$, where $p_1, ..., p_k$ are distinct primes, and $\alpha_i \in \mathbb{N}$ for all $1 \leq i \leq k$.

For all such n, $\sigma(n) = (\frac{p_1^{\alpha_1+1}-1}{p_1-1})...(\frac{p_k^{\alpha_k+1}-1}{p_k-1})$. We will call this *Property 2*, and refer back to it

throughout the rest of this problem.

Claim: $\sigma(175) = 248$.

Proof. $175 \geq 2$, and we can write $175 = p_1^{\alpha_1}p_2^{\alpha_2} = 5^2 7^1$, $\implies p_1 = 5$, $\alpha_1 = 2$, $p_2 = 7$, and $\alpha_2 = 1$.

By *Property 2*, $\implies \sigma(175) = (\frac{p_1^{\alpha_1+1}-1}{p_1-1})...(\frac{p_k^{\alpha_k+1}-1}{p_k-1}) = (\frac{5^{2+1}-1}{5-1})(\frac{7^{1+1}-1}{7-1}) = (\frac{124}{4})(\frac{48}{6}) = 31(8) = 248$,

which is exactly what we want to show, and thus concludes the proof. □

Claim: $\sigma(231) = 384$.

Proof. $231 \geq 2$, and we can write $231 = p_1^{\alpha_1}p_2^{\alpha_2}p_3^{\alpha_3} = 3^1 7^1 11^1$, $\implies p_1 = 3$, $\alpha_1 = 1$, $p_2 = 7$, $\alpha_2 = 1$, $p_3 = 11$, $\alpha_3 = 1$.

By *Property 2*, $\implies \sigma(231) = (\frac{p_1^{\alpha_1+1}-1}{p_1-1})...(\frac{p_k^{\alpha_k+1}-1}{p_k-1}) = (\frac{3^{1+1}-1}{3-1})(\frac{7^{1+1}-1}{7-1})(\frac{11^{1+1}-1}{11-1}) = (\frac{8}{2})(\frac{48}{6})(\frac{120}{10}) = 4(8)(12) = 384$,

which is exactly what we want to show, and thus concludes the proof. □

Claim: The sum of the *even* positive divisors of 10000 is 23430.

Proof. Since $10000 = 2^4 5^4$, all positive divisors of 10000 are given by $d = 2^{\alpha_1}5^{\alpha_2}$, where $0 \leq \alpha_1, \alpha_2 \leq 4$.

For a positive divisor of 10000 to be even, 2 must be one of its factors,

so we know all even positive divisors of 10000 are given by

$$d = 2^{\alpha_1}5^{\alpha_2}, \text{ where } 1 \leq \alpha_1 \leq 4, \text{ and } 0 \leq \alpha_2 \leq 4$$

To find the sum of all such even positive divisors d, we calculate

$$\sum_{\substack{1\leq\alpha_1\leq 4\\ 0\leq\alpha_2\leq 4}} 2^{\alpha_1}5^{\alpha_2}$$

Due to the distributive property of addition, we can split this sum as follows:

$$\sum_{\substack{1\leq\alpha_1\leq 4\\ 0\leq\alpha_2\leq 4}} 2^{\alpha_1}5^{\alpha_2} = \sum_{1\leq\alpha_1\leq 4} 2^{\alpha_1} \sum_{0\leq\alpha_2\leq 4} 5^{\alpha_2}$$

Now, we can easily compute that

$$\sum_{\substack{1\leq\alpha_1\leq 4\\ 0\leq\alpha_2\leq 4}} 2^{\alpha_1}5^{\alpha_2} = (2^1+2^2+2^3+2^4)(5^0+5^1+5^2+5^3+5^4) = (2+4+8+16)(1+5+25+125+625) = 30(781) = 23430$$

which is exactly what we want to show, and thus concludes the proof □

Problem 4

(5 points). Show that for each $n \in \mathbb{N}$,

$$\textit{(Product 1)} := \prod_{d|n, d\in\mathbb{N}} d = n^{\frac{\tau(n)}{2}}$$

Proof. Note: if $n = 1$, then n's only positive divisor is 1, so $\tau(n) = 1$, and

$$\prod_{d|n, d\in\mathbb{N}} d = 1 = 1^{\frac{1}{2}} = \sqrt{1} = 1$$

so we just need to show that

$$\prod_{d|n, d\in\mathbb{N}} d = n^{\frac{\tau(n)}{2}}$$

for all $n \in \mathbb{N}, n \geq 2$.

Note that we can write any such n as $n = p_1^{\alpha_1}...p_k^{\alpha_k}$, where $p_1, ..., p_k$ are distinct primes and $\alpha_i \in \mathbb{N}$ for all $1 \leq i \leq k$.

Therefore, all positive divisors, d, of such n, are given by $d = p_1^{\beta_1}...p_k^{\beta_k}$, where $0 \leq \beta_i \leq \alpha_i$ for all $1 \leq i \leq k$.

$$\implies \textit{(Product 1)} := \prod_{d|n, d\in\mathbb{N}} d = \prod_{\substack{0\leq\beta_1\leq\alpha_1 \\ \vdots \\ 0\leq\beta_k\leq\alpha_k}} p_1^{\beta_1}...p_k^{\beta_k}$$

Since $p_1, ..., p_k$ are distinct primes, we can consider their contributions to *Product 1* separately.

For each p_i, we can rewrite all positive divisors d as

$$d = p_i^{\beta_i} \prod_{\substack{1\leq j\leq k \\ j\neq i}} p_j^{\beta_j},\ 0 \leq \beta_i \leq \alpha_i,\ 0 \leq \beta_j \leq \alpha_j$$

We can select each β_j in $(\alpha_j + 1)$ ways, and we can make each selection independently, so, for a given β_i, there are

$$\prod_{\substack{1\leq j\leq k \\ j\neq i}} (\alpha_j + 1)$$

positive divisors d with $p_i^{\beta_i}$ as the largest power of p_i that divides them,

and each such d contributes β_i copies of p_i to *Product 1*, so the product of all such divisors d contributes exactly

$$\beta_i \prod_{\substack{1\leq j\leq k \\ j\neq i}} (\alpha_j + 1)$$

copies of p_i to *Product 1*.

This is true for all $0 \leq \beta_i \leq \alpha_i$, so we know there are exactly

$$\sum_{0\leq\beta_i\leq\alpha_i} (\beta_i \prod_{\substack{1\leq j\leq k \\ j\neq i}} (\alpha_j + 1)) = (\sum_{0\leq\beta_i\leq\alpha_i} \beta_i) \prod_{\substack{1\leq j\leq k \\ j\neq i}} (\alpha_j + 1) = (\frac{\alpha_i(\alpha_i + 1)}{2}) \prod_{\substack{1\leq j\leq k \\ j\neq i}} (\alpha_j + 1) = \frac{\alpha_i}{2} \prod_{1\leq j\leq k} (\alpha_j + 1)$$

copies of p_i in *Product 1.*

This is true for all p_i (for all $1 \leq i \leq k$), so we can now calculate that

$$\textit{(Product 1)} := \prod_{d|n, d\in\mathbb{N}} d = p_1^{\frac{\alpha_1}{2}\prod_{1\leq j\leq k}(\alpha_j+1)} ... p_k^{\frac{\alpha_k}{2}\prod_{1\leq j\leq k}(\alpha_j+1)} = p_1^{\frac{\alpha_1}{2}(\alpha_1+1)...(\alpha_k+1)} ... p_k^{\frac{\alpha_k}{2}(\alpha_1+1)...(\alpha_k+1)}$$

By *Property 1* of **Problem 2**, we know $\tau(n) = (\alpha_1 + 1)...(\alpha_k + 1)$.

$$\implies \textit{(Product 1)} := \prod_{d|n, d\in\mathbb{N}} d = p_1^{\frac{\alpha_1}{2}\tau(n)} ... p_k^{\frac{\alpha_k}{2}\tau(n)} = (p_1^{\alpha_1} ... p_k^{\alpha_k})^{\frac{\tau(n)}{2}} = n^{\frac{\tau(n)}{2}},$$

which is exactly what we want to show, and thus concludes the proof. □

Problem 5

(Very difficult Bonus, 5 points). On planet E, there is an advanced alien civilization. There is a group of aliens who would like to go on a field trip to Planet Earth, their zoo. Some of them are friends, some are strangers. They will take two spaceships for their journey. However, on planet E, they always insist that on their spaceships, every alien should be friends with an even number of other aliens (hence the name E). Show that the number of ways the aliens may be assigned to the two spaceships is of the form 2^k, $k \in N$.

Proof. Note that we can model the set of aliens and their friendships via a graph $G = (A, F)$,

where $A :=$ {all aliens on planet E}, and $F = \{(a_i, a_j) \in A \text{ x } A \mid a_i \text{ and } a_j \text{ are friends}\}$.

Also Note: *good configuration* refers to any grouping of aliens that results in all aliens

having an even number of friends on their spaceship.

First, we must show that, for any G, there exists at least one good configuration.

To do so, we will induct on the number of aliens, which we will call n.

Base Case: $n = 1$. There is only one alien, who has no friends, so he can go into Spaceship 1,

and we have a good configuration.

Inductive Hypothesis: Assume at least one good configuration exists with n aliens,

regardless of the set of friendships, for all $1 \leq n \leq m$.

Inductive Step: We want to show that at least one good configuration exists with $m + 1$ aliens,

regardless of the set of friendships.

Consider an arbitrary graph $G = (A, F)$ with $|A| = m + 1$.

Note: If all aliens have an even number of friends, we can put all aliens in the same Spaceship,

and we will have a good configuration.

Thus, we just need to show that we can always find a good configuration if at least one of the

$m+1$ aliens has an odd number of friends.

Take one such alien with an odd number of friends, a_i and remove it from the graph.

For each deleted friend $(a_i, a_j) \in F$, create a new friend $(a_x, a_j) \in F$, that is,

direct all friendships with a_i towards a distinct alien a_x.

Now, we have a graph with m aliens, so the **Inductive Hypothesis** guarantees

we can find a good configuration for the graph, which we'll call C.

Since a_i had an odd number of friends, we know that C places an odd number

of a_i's friends in one Spaceship and an even number of a_i's friends in the other Spaceship.

Since a_x is friends with all of a_i's friends, we know C

places a_x in the Spaceship that has an even number of a_i's friends, as these friends contribute an even number to the

total number of friends a_x has in its Spaceship, which is guaranteed to be even.

Now, add a_i back to the graph, specifically placing it in the same Spaceship as a_x.

Also, remove all the friendships $(a_x, a_j) \in F$ that were created after deleting a_i, to create a new configuration C'.

Since a_i has an even number of friends in its Spaceship, C' works for a_i.

Since a_i's friends in a_i's Spaceship both lost and gained a friend,

they still have even numbers of friends on their Spaceship.

All other aliens have the same number of friends on their Spaceship C' as in C, so they are all guaranteed

to have an even number of friends on their Spaceship.

Thus, C' guarantees that all aliens have an even number of friends on their Spaceship, so C' is a good configuration,

which is exactly what we want to show, and concludes the proof of existence.

Now, we want to show that the number of possible good configurations is always of the form 2^k for some $k \in \mathbb{N}$.

Note: If we have one good configuration, X, we can consider operations $op_i \in (\mathbb{Z}/2\mathbb{Z})^n$

which send X to another good configuration.

Also note: All good configurations can be reached with one operation, so the number of good configurations

equals the size of the set $S := \{\text{all operations that send } X \text{ to a good configuration}\}$.

If we let each operation $op_i = (o_1, ..., o_n)$, $o_i \in (\mathbb{Z}/2\mathbb{Z})$ then each o_i indicates an operation to perform on the

corresponding alien a_i, where 0 indicates leaving a_i in its Spaceship and 1 indicates

moving a_i to the opposite Spaceship.

Thus, we know $(1, 1, ..., 1) \in S$, as taking a good configuration and flipping all aliens to opposite Spaceships

always results in a symmetrical good configuration.

If we can show that S forms a vector space over $(\mathbb{Z}/2\mathbb{Z})$, then, since $\{(1, 1, ..., 1)\} \subseteq S$, $dim[S] \geq 1$,

$\implies$ S is isomorphic to $(\mathbb{Z}/2\mathbb{Z})^k$ for some $k \in \mathbb{N}$,

$\implies$ number of good configurations $= |S| = 2^k$ for some $k \in \mathbb{N}$, since $(\mathbb{Z}/2\mathbb{Z})^k$ has 2^k elements.

Consider $S' :=$ {all operations that send all good configurations to a good configuration}.

For all $u, v \in S'$, applying $(u + v)$ to any good configuration X is identical to first applying u to X to form a good

configuration X', then applying v to X', which results in a good configuration X".

Therefore, $(u + v)$ sends any good configuration to some good configuration,

so $(u + v) \in S'$, so S' is closed under addition.

Similarly, for all $v \in S'$, $0(v) = (0, ..., 0) \in S'$ since doing nothing to any good configuration always results

in a good configuration.

Also, trivially, for all $v \in S'$, $1(v) = v \in S'$, so S' is closed under scalar multiplication.

Therefore, S' is a vector space over $(\mathbb{Z}/2\mathbb{Z})$ with dimension ≥ 1, so S' has 2^k elements for some $k \in \mathbb{N}$.

Thus, it suffices to show that $S' = S$. We can do this by considering $op_1 \in S$ such that op_1 sends a good

configuration X to a good configuration Y. We can then define

S" := {all operations that send Y to a good configuration}, and let $op_2 \in S$" such that.

op_2 sends Y to a good configuration Z. Since Z is a good configuration, we know $\exists$ $op_3 \in S$

such that op_3 sends X to Z.

Note that $op_1 + op_2 = op_3 \implies op_3 + op_1 = op_2$, since $a + b \equiv a - b \pmod 2$, so it suffices to show that $op_2 \in S$.

Assume to the contrary that $op_2 \notin S$. Then applying $(op_1 + op_2) = (op_2 + op_1)$ to X is identical to first applying op_2

then applying op_1 to X. Since $op_2 \notin S$, we know applying op_2 to X results in a bad configuration, X'. If we then apply

op_1 to X', this is identical to flipping aliens as needed to get X from X', then applying op_1 to X to get a good

configuration Y, then flipping the aliens back to get a bad configuration X". Thus, applying $(op_1 + op_2) = (op_2 + op_1)$

to X results in a bad configuration. However, we know that this is identical to first applying op_1 to X

to get a good configuration Y, then applying op_2 to Y to get a good configuration Z. Thus, we have

X" (bad configuration) = Z (good configuration), which is a contradiction, so we know $op_2 \in S$.

Therefore, any operation that sends a fixed good configuration to a good configuration also sends all good configurations to a good configuration.

Therefore, $S = S'$, so S is a vector space over $(\mathbb{Z}/2\mathbb{Z})$ with dimension ≥ 1, so

number of good configurations $= |S| = |S'| = 2^k$ for some $k \in \mathbb{N}$,

which is exactly what we want to show, and thus concludes the proof. □

Assignment 10

Problem 1

(5 points). Prove that for every $n \in \mathbb{N}$,

$$n = \sum_{d|n} \mu(\frac{n}{d})\sigma(d).$$

Proof. By the Möbius Inversion Formula, if $f : \mathbb{N} \to \mathbb{C}$, and for all $n \in \mathbb{N}$,

$$g(n) := \sum_{d|n} f(d), \text{ then}$$

$$f(n) = \sum_{d|n} \mu(\frac{n}{d})g(d), \text{ where } \mu \text{ is the möbius function } \mu(n) = \begin{cases} 1 \text{ if } n = 1 \\ (-1)^r \text{ if } n = p_1...p_r, \ p_i \text{ distinct primes} \\ 0 \text{ otherwise} \end{cases}$$

Note: $\sigma(n) = \sum_{d|n} d = \sum_{d|n} f(d)$, where $f(d) = d$ for all $d \in \mathbb{N}$. Thus, by the Möbius Inversion Formula,

$$f(n) = n = \sum_{d|n} \mu(\frac{n}{d})\sigma(d), \text{ for all } n \in \mathbb{N}$$

which is exactly what we want to show, and thus concludes the proof. □

Problem 2

(5 points). Define the function

$$\theta(n) := \begin{cases} log(p) \text{ if } n = p^\alpha, \ p \text{ prime}, \ a \geq 1 \\ 0 \text{ otherwise.} \end{cases}$$

Using the Möbius Inversion Formula, prove that

$$\theta(n) = \sum_{d|n} \mu(\frac{n}{d})log(d)$$

Proof. By the Möbius Inversion Formula, it suffices to show that

$$g(n) = log(n) = \sum_{d|n} \theta(d) \tag{27}$$

Note: $g(1) = log(1) = 0 = \sum_{d|1} \theta(d) = \theta(1) = 0$, so we just need to show that (1) holds for all $n \in \mathbb{N}$ such that $n \geq 2$.

For all such $n \geq 2$, we can write the prime factorization $n = p_1^{\alpha_1}...p_k^{\alpha_k}$, where $p_1, ..., p_k$ are distinct primes,

and $\alpha_i \in \mathbb{N}$ for all $1 \leq i \leq k$.

Therefore, all positive divisors are of the form $d = p_1^{\beta_1}...p_k^{\beta_k}$, where $0 \leq \beta_i \leq \alpha_i$ for all $1 \leq i \leq k$.

$$\implies \sum_{d|n} \theta(d) = \sum_{\substack{0 \leq \beta_1 \leq \alpha_1 \\ \vdots \\ 0 \leq \beta_k \leq \alpha_k}} \theta(p_1^{\beta_1}...p_k^{\beta_k})$$

Note: By definition, $\theta(p_1^{\beta_1}...p_k^{\beta_k}) = 0$ unless $\beta_i \geq 1$, $\beta_j = 0$ for all $1 \leq j \leq k$ such that $j \neq i$.

$$\implies \sum_{d|n} \theta(d) = \sum_{\substack{0 \leq \beta_1 \leq \alpha_1 \\ \vdots \\ 0 \leq \beta_k \leq \alpha_k}} \theta(p_1^{\beta_1}...p_k^{\beta_k}) = \sum_{\substack{1 \leq i \leq k \\ 1 \leq \beta_i \leq \alpha_i}} \theta(p_i^{\beta_i})$$

For each $p_i^{\beta_i}$, $\beta_i \geq 1$, $\theta(p_i^{\beta_i}) = log(p_i)$.

For each p_i, β_i ranges from 1 to α_i, so each p_i contributes exactly α_i copies of $log(p_i)$ to the summation.

$$\begin{aligned} \implies \sum_{d|n} \theta(d) = \sum_{\substack{1 \leq i \leq k \\ 1 \leq \beta_i \leq \alpha_i}} \theta(p_i^{\beta_i}) = & \underbrace{(log(p_1) + ... + log(p_1))}_{\alpha_1 \text{ times}} + \underbrace{(log(p_2) + ... + log(p_2))}_{\alpha_2 \text{ times}} + ... + \underbrace{(log(p_k) + ... + log(p_k))}_{\alpha_k \text{ times}} \\ = & \ \alpha_1 log(p_1) + \alpha_2 log(p_2) + ... + \alpha_k log(p_k) = log(p_1^{\alpha_1}) + log(p_2^{\alpha_2}) + ... + log(p_k^{\alpha_k}) \\ = & \ log(p_1^{\alpha_1}...p_k^{\alpha_k}) = log(n) = g(n) \end{aligned}$$

Therefore, (1) holds for all $n \in \mathbb{N}$.

Therefore, by the Möbius Inversion Formula,

$$\theta(n) = \sum_{d|n} \mu(\frac{n}{d}) log(d) \text{ for all } n \in \mathbb{N},$$

which is exactly what we want to show, and thus concludes the proof. □

Problem 3

(5 points). Find a formula for

$$\sum_{d|n} \mu(d)\varphi(d)$$

in terms of the prime factorization $n = p_1^{\alpha_1}...p_k^{\alpha_k}$, p_i distinct primes. Hint: $n \to \mu(n)\varphi(n)$is a multiplicative function.

Claim: Define $g(n) := \sum_{d|n} \mu(d)\varphi(d)$. We claim that

$$g(n) = \sum_{d|n} \mu(d)\varphi(d) = \begin{cases} 1 \text{ if } n = 1 \\ (2 - p_1)...(2 - p_k) \text{ otherwise} \end{cases} \tag{28}$$

Proof. Case 1: $n = 1$, so the only positive divisor of n is 1. We know $\mu(1) := 1$ and $\varphi(1) = 1$, so

$$g(1) = \sum_{d|1} \mu(d)\varphi(d) = \mu(1)\varphi(1) = 1(1) = 1, \text{ as required.}$$

Case 2: $n > 1$. Now, we can impose the restriction $\alpha_i \in \mathbb{N}$ for all $1 \leq i \leq k$ on the prime factorization $n = p_1^{\alpha_1}...p_k^{\alpha_k}$.

We know all positive divisors of n are given by $d = p_1^{\beta_1}...p_k^{\beta_k}$, where $\beta_i \in \mathbb{Z}$, and $0 \leq \beta_i \leq \alpha_i$ for all $1 \leq i \leq k$.

$$\implies g(n) = \sum_{d|n} \mu(d)\varphi(d) = \sum_{\substack{0\leq\beta_1\leq\alpha_1 \\ \vdots \\ 0\leq\beta_k\leq\alpha_k}} \mu(p_1^{\beta_1}...p_k^{\beta_k})\varphi(p_1^{\beta_1}...p_k^{\beta_k})$$

Since $p_1, ..., p_k$ are distinct primes, and we want to find $g(n)$ in terms of $p_1, ..., p_k$,

we want to consider the contributions of each prime p_i separately.

Since $n \to \mu(n)\varphi(n)$ is a multiplicative function, we know that

$$\mu(p_1^{\beta_1}...p_k^{\beta_k})\varphi(p_1^{\beta_1}...p_k^{\beta_k}) = \mu(p_1^{\beta_1})\varphi(p_1^{\beta_1})...\mu(p_k^{\beta_k})\varphi(p_k^{\beta_k})$$

$$\implies g(n) = \sum_{\substack{0\leq\beta_1\leq\alpha_1 \\ \vdots \\ 0\leq\beta_k\leq\alpha_k}} \mu(p_1^{\beta_1}...p_k^{\beta_k})\varphi(p_1^{\beta_1}...p_k^{\beta_k}) = \sum_{\substack{0\leq\beta_1\leq\alpha_1 \\ \vdots \\ 0\leq\beta_k\leq\alpha_k}} \mu(p_1^{\beta_1})\varphi(p_1^{\beta_1})...\mu(p_k^{\beta_k})\varphi(p_k^{\beta_k})$$

By the distributive property of arithmetic addition, we know that

$$g(n) = \sum_{\substack{0\leq\beta_1\leq\alpha_1 \\ \vdots \\ 0\leq\beta_k\leq\alpha_k}} \mu(p_1^{\beta_1})\varphi(p_1^{\beta_1})...\mu(p_k^{\beta_k})\varphi(p_k^{\beta_k}) = (\sum_{0\leq\beta_1\leq\alpha_1} \mu(p_1^{\beta_1})\varphi(p_1^{\beta_1}))...(\sum_{0\leq\beta_k\leq\alpha_k} \mu(p_k^{\beta_k})\varphi(p_k^{\beta_k}))$$

For each p_i, $1 \leq i \leq k$,

$$\begin{aligned}
\sum_{0\leq\beta_i\leq\alpha_i} \mu(p_i^{\beta_i})\varphi(p_i^{\beta_i}) = \quad & \mu(p_i^0)\varphi(p_i^0) + \mu(p_i^1)\varphi(p_i^1) + \mu(p_i^2)\varphi(p_i^2) + ... + \mu(p_i^{\alpha_i})\varphi(p_i^{\alpha_i}) \\
= \quad & \mu(1)\varphi(1) + \mu(p_i)\varphi(p_i) + \mu(p_i^2)\varphi(p_i^2) + ... + \mu(p_i^{\alpha_i})\varphi(p_i^{\alpha_i}) \\
= \quad & 1(1) + (-1)(p_i - 1) + (0)\varphi(p_i^2) + ... + (0)\varphi(p_i^{\alpha_i}) = 1 + (1 - p_i) = 2 - p_i,
\end{aligned}$$

since $\mu(p_i^{\beta_i}) = 0$ for all $\beta_i \geq 2$. This is true for all p_i, so we know that, for all $n \geq 2$,

$$g(n) = \sum_{d|n} \mu(d)\varphi(d) = \sum_{\substack{0\leq\beta_1\leq\alpha_1 \\ \vdots \\ 0\leq\beta_k\leq\alpha_k}} \mu(p_1^{\beta_1}...p_k^{\beta_k})\varphi(p_1^{\beta_1}...p_k^{\beta_k}) = (2 - p_1)...(2 - p_k) \text{ as required.}$$

Therefore, we have shown that

$$g(n) = \sum_{d|n} \mu(d)\varphi(d) = \begin{cases} 1 \text{ if } n = 1 \\ (2 - p_1)...(2 - p_k) \text{ otherwise} \end{cases},$$

which is exactly what we want to show, and thus concludes the proof. □

Problem 4

(5 points). Prove that for every $n \in \mathbb{N}$,

$$\sum_{k=1}^{n} gcd(k,n) = \sum_{d|n} d\varphi(\frac{n}{d}) = n\sum_{d|n} \frac{\varphi(d)}{d}$$

Hint: See the proof of Gauss's Lemma.

Proof. Let $D := \{$all positive divisors of n$\}$. For all $k \in \{1, 2, ..., n\}$, $gcd(k,n) \in D$.

Therefore, every term in

$$\sum_{k=1}^{n} gcd(k,n)$$

is an element of D. Now, let's count how many times each $d \in D$ appears as a term in the sum.

For each $k \in \{1, 2, ..., n\}$, let $gcd(k,n) = d \in D$. Since $d|k$, we know $k = id \implies i = \frac{k}{d}$ for some $i \in \mathbb{N}$.

Since $gcd(k,n) = d$, we know $gcd(\frac{k}{d}, \frac{n}{d}) = gcd(i, \frac{n}{d}) = 1$. Since $1 \leq k \leq n$ and $i \in \mathbb{N}$, we also know $1 \leq i \leq \frac{n}{d}$.

By the definition of the Euler Totient, we know that, for each $d \in D$, there are exactly

$\varphi(\frac{n}{d}) := \#\{1 \leq i \leq \frac{n}{d} | gcd(i, \frac{n}{d}) = 1\}$ possibilities for i.

Therefore, each $d \in D$ appears exactly $\varphi(\frac{n}{d})$ times in

$$\sum_{k=1}^{n} gcd(k,n).$$

Therefore, we know that

$$\sum_{k=1}^{n} gcd(k,n) = \sum_{d|n} d\varphi(\frac{n}{d}), \text{ which proves the left equality.}$$

For the right equality, note that for each $d \in D$, $\frac{n}{d} \in D$, and $\frac{n}{\frac{n}{d}} = d$, so we know that

$$\sum_{d|n} d\varphi(\frac{n}{d}) = \sum_{d|n} \frac{n}{d}\varphi(d) = n\sum_{d|n} \frac{\varphi(d)}{d}, \text{ which proves the right equality.}$$

Thus, we have shown that

$$\sum_{k=1}^{n} gcd(k,n) = \sum_{d|n} d\varphi(\frac{n}{d}) = n\sum_{d|n} \frac{\varphi(d)}{d},$$

which is exactly what we want to show, and thus concludes the proof. □

Problem 5

(5 points). Find a formula for $\sum_{d|n} \frac{\varphi(d)}{d}$ in terms of the prime factorization $n = p_1^{\alpha_1}...p_k^{\alpha_k}$, p_i distinct primes.

Use the previous exercise to find a formula for

$$\sum_{k=1}^{n} gcd(k,n)$$

in terms of the prime factorization of n.

Claim:

$$\sum_{d|n} \frac{\varphi(d)}{d} = \begin{cases} 1 \text{ if } n = 1 \\ \prod_{i=1}^{k}(1 + \alpha_i(1 - \frac{1}{p_i})) \text{ if } n > 1 \end{cases}$$

Proof. Case 1: $n = 1$, so 1 is n's only positive divisor, so

$$\sum_{d|n} \frac{\varphi(d)}{d} = \frac{1}{1} = 1,$$

so we just need to show our formula holds for $n > 1$.

Case 2: $n > 1$, so we can impose the restriction $\alpha_i \in \mathbb{N}$ for all $1 \leq i \leq k$ on the prime factorization $n = p_1^{\alpha_1}...p_k^{\alpha_k}$.

Note: all positive divisors $d|n$ are given by $d = p_1^{\beta_1}...p_k^{\beta_k}$, where $0 \leq \beta_i \leq \alpha_i$ for all $1 \leq i \leq k$.

$$\implies \sum_{d|n} \frac{\varphi(d)}{d} = \sum_{\substack{0 \leq \beta_1 \leq \alpha_1 \\ \vdots \\ 0 \leq \beta_k \leq \alpha_k}} \frac{\varphi(p_1^{\beta_1}...p_k^{\beta_k})}{p_1^{\beta_1}...p_k^{\beta_k}}$$

By **Problem 3** of Homework 8, we know $\varphi(n)$ is a multiplicative function.

$$\implies \sum_{d|n} \frac{\varphi(d)}{d} = \sum_{\substack{0 \leq \beta_1 \leq \alpha_1 \\ \vdots \\ 0 \leq \beta_k \leq \alpha_k}} \frac{\varphi(p_1^{\beta_1}...p_k^{\beta_k})}{p_1^{\beta_1}...p_k^{\beta_k}} = \sum_{\substack{0 \leq \beta_1 \leq \alpha_1 \\ \vdots \\ 0 \leq \beta_k \leq \alpha_k}} \frac{\varphi(p_1^{\beta_1})...\varphi(p_k^{\beta_k})}{p_1^{\beta_1}...p_k^{\beta_k}}$$

By the distributive property of arithmetic addition, we know that

$$\sum_{\substack{0 \leq \beta_1 \leq \alpha_1 \\ \vdots \\ 0 \leq \beta_k \leq \alpha_k}} \frac{\varphi(p_1^{\beta_1})...\varphi(p_k^{\beta_k})}{p_1^{\beta_1}...p_k^{\beta_k}} = (\sum_{0 \leq \beta_1 \leq \alpha_1} \frac{\varphi(p_1^{\beta_1})}{p_1^{\beta_1}})...(\sum_{0 \leq \beta_k \leq \alpha_k} \frac{\varphi(p_k^{\beta_k})}{p_k^{\beta_k}})$$

For all $1 \leq i \leq k$, we know that

$$\frac{\varphi(p_i^0)}{p_i^0} = \frac{\varphi(1)}{1} = \frac{1}{1} = 1$$

Also, for all $1 \leq i \leq k$, $\beta_i \geq 1$, we know that

$$\frac{\varphi(p_i^{\beta_i})}{p_i^{\beta_i}} = \frac{p_i^{\beta_i}(1 - \frac{1}{p_i})}{p_i^{\beta_i}} = 1 - \frac{1}{p_i}$$

Therefore, we can easily calculate that, for all $1 \leq i \leq k$,

$$\sum_{0 \leq \beta_i \leq \alpha_i} \frac{\varphi(p_i^{\beta_i})}{p_i^{\beta_i}} = \frac{\varphi(p_i^0)}{p_i^0} + \frac{\varphi(p_i^1)}{p_i^1} + \frac{\varphi(p_i^2)}{p_i^2} + ... + \frac{\varphi(p_i^{\alpha_i})}{p_i^{\alpha_i}} = 1 + \underbrace{(1 - \frac{1}{p_i}) + ... + (1 - \frac{1}{p_i})}_{\alpha_i \text{ times}} = 1 + \alpha_i(1 - \frac{1}{p_i})$$

This is true for all p_i, so we know that

$$(\sum_{0 \leq \beta_1 \leq \alpha_1} \frac{\varphi(p_1^{\beta_1})}{p_1^{\beta_1}})...(\sum_{0 \leq \beta_k \leq \alpha_k} \frac{\varphi(p_k^{\beta_k})}{p_k^{\beta_k}}) = \prod_{i=1}^{k}(1 + \alpha_i(1 - \frac{1}{p_i}))$$

for all $n > 1$. Thus, we have shown that

$$\sum_{d|n} \frac{\varphi(d)}{d} = \begin{cases} 1 \text{ if } n = 1 \\ \prod_{i=1}^{k}(1 + \alpha_i(1 - \frac{1}{p_i})) \text{ if } n > 1 \end{cases}$$

which is exactly what we want to show, and thus concludes the proof. □

Claim:

$$\sum_{k=1}^{n} gcd(k, n) = \begin{cases} 1 \text{ if } n = 1 \\ p_1^{\alpha_1}...p_k^{\alpha_k} \prod_{i=1}^{k}(1 + \alpha_i(1 - \frac{1}{p_i})) \text{ if } n > 1 \end{cases}$$

Proof. By **Problem 4**, we know

$$\sum_{k=1}^{n} gcd(k, n) = n \sum_{d|n} \frac{\varphi(d)}{d}$$

By the previous proof, we know

$$\sum_{d|n} \frac{\varphi(d)}{d} = \begin{cases} 1 \text{ if } n = 1 \\ \prod_{i=1}^{k}(1 + \alpha_i(1 - \frac{1}{p_i})) \text{ if } n > 1 \end{cases}$$

Therefore, we know

$$\sum_{k=1}^{n} gcd(k, n) = n \sum_{d|n} \frac{\varphi(d)}{d} = \begin{cases} 1(1) = 1 \text{ if } n = 1 \\ p_1^{\alpha_1}...p_k^{\alpha_k} \prod_{i=1}^{k}(1 + \alpha_i(1 - \frac{1}{p_i})) \text{ if } n > 1 \end{cases}$$

which is exactly what we want to show, and thus concludes the proof. □

Problem 6

(Bonus, 5 points). Define

$$p_n(x) := \sum_{d|n} \mu(d) x^{\frac{n}{d}} \in \mathbb{Z}[x] \text{ for each } n \in \mathbb{N}.$$

Let $p \in \mathbb{Z}[x]$ be a polynomial with leading coefficient 1 and degree k with roots $z_1, ..., z_k \in \mathbb{C}$.

Show that for each integer $\ell \geq 2$ and each $n \in \mathbb{N}$,

$$n | \sum_{d|n} \sum_{j=1}^{k} z_j^{\ell-1} p_d'(z_j^{\ell})$$

makes sense and is true. Make sense of this even when $\ell = 1$, and prove it. The prime denotes differentiation.

Proof. First, note that

$$p_n(x) := \sum_{d|n} \mu(d) x^{\frac{n}{d}} = \sum_{d|n} \mu(\frac{n}{d}) x^d \text{ for each } n \in \mathbb{N}$$

Therefore,

$$p_n'(x) = \frac{d}{dx}(\sum_{d|n} \mu(\frac{n}{d}) x^d) = \sum_{d|n} \mu(\frac{n}{d}) \frac{d}{dx}(x^d) = \sum_{d|n} \mu(\frac{n}{d}) d(x)^{d-1} \text{ for each } n \in \mathbb{N}$$

Now, for each root, $z_1, ..., z_k \in \mathbb{C}$, let's define a function $p_{z_j} : \mathbb{N} \to \mathbb{C}$, where

$$p_{z_j}(n) := n(z_j^{\ell})^{n-1} \text{ for each } n \in \mathbb{N}$$

Now, let's consider the right hand side of the divisibility statement.

Case 1: $\ell \geq 2$. Note that, for all $\ell \geq 2$,

$$\sum_{d|n}\sum_{j=1}^{k} z_j^{\ell-1} p_d'(z_j^\ell) = \sum_{d|n}\sum_{j=1}^{k}(z_j^{\ell-1}\sum_{q|d}\mu(\frac{d}{q})q(z_j^\ell)^{q-1}) = \sum_{d|n}\sum_{j=1}^{k}(z_j^{\ell-1}\sum_{q|d}\mu(\frac{d}{q})p_{z_j}(q)) = \sum_{j=1}^{k} z_j^{\ell-1}\sum_{d|n}\sum_{q|d}\mu(\frac{d}{q})p_{z_j}(q)$$

Let $f_{z_j} : \mathbb{N} \to \mathbb{C}$ be defined by

$$f_{z_j}(d) = \sum_{q|d}\mu(\frac{d}{q})p_{z_j}(q) \text{ for each } d \in \mathbb{N}$$

Since $p_{z_j} : \mathbb{N} \to \mathbb{C}$ for all $1 \leq j \leq k$, the Möbius Inversion Formula guarantees that

$$p_{z_j}(n) := n(z_j^\ell)^{n-1} = \sum_{d|n} f_{z_j}(d)$$

Applying this to the right hand side of the divisibility statement, we find that

$$\begin{aligned}\sum_{d|n}\sum_{j=1}^{k} z_j^{\ell-1} p_d'(z_j^\ell) = &\quad \sum_{j=1}^{k} z_j^{\ell-1}\sum_{d|n}\sum_{q|d}\mu(\frac{d}{q})p_{z_j}(q) = \sum_{j=1}^{k} z_j^{\ell-1}\sum_{d|n} f_{z_j}(d) = \sum_{j=1}^{k} z_j^{\ell-1} p_{z_j}(n) = \sum_{j=1}^{k} z_j^{\ell-1} n(z_j^\ell)^{n-1} \\ = &\quad n\sum_{j=1}^{k} z_j^{\ell-1} z_j^{\ell n-\ell} = n\sum_{j=1}^{k} z_j^{\ell n-\ell+\ell-1} = n\sum_{j=1}^{k} z_j^{\ell n-1}\end{aligned}$$

Note that

$$n|\sum_{d|n}\sum_{j=1}^{k} z_j^{\ell-1} p_d'(z_j^\ell) = n\sum_{j=1}^{k} z_j^{\ell n-1} \iff \sum_{j=1}^{k} z_j^{\ell n-1} = (z_1^{\ell n-1} + ... + z_k^{\ell n-1}) \in \mathbb{Z},$$

so it suffices to show that

$$\sum_{j=1}^{k} z_j^{\ell n-1} = (z_1^{\ell n-1} + ... + z_k^{\ell n-1}) \in \mathbb{Z}$$

for all $\ell \geq 2, n \in \mathbb{N}$. Since $\ell \geq 2$, we know $\ell n \geq 2 \implies \ell n - 1 \geq 1 \implies \ell n - 1 \in \mathbb{N}$. Thus, it suffices to show that

$$s_h := \sum_{j=1}^{k} z_j^h \in \mathbb{Z} \text{ for all } h \in \mathbb{N}$$

Since the polynomial p has k roots and leading coefficient 1, we can write

$$p(x) = a_k x^k + a_{k-1}x^{k-1} + ... + a_1 x^1 + a_0 x^0 = x^k + a_{k-1}x^{k-1} + ... + a_1 x + a_0$$

Since $p(x) \in \mathbb{Z}[x]$, we know that $a_i \in \mathbb{Z}$ for all $0 \leq i \leq k$. Using Newton's Identities, we can define

$e_i :=$ the sum of all distinct products of i roots of $p(x)$ for all $1 \leq i \leq k$. By Vieta's Formulae, we know

$$e_i = (-1)^i \frac{a_{k-i}}{a_k} = (-1)^i a_{k-i} \text{ for all } 1 \leq i \leq k$$

Since $(-1)^i, a_{k-i} \in \mathbb{Z}$ for all $1 \leq i \leq k$, we know $e_i = (-1)^i a_{k-i} \in \mathbb{Z}$ for all $1 \leq i \leq k$.

Also, Newton's Identities tell us that, for all $1 \leq h \leq k$, $h \in \mathbb{N}$,

$$s_h := \sum_{j=1}^{k} z_j^h = (-1)^{h-1} h e_h + \sum_{i=1}^{h-1}(-1)^{h-1+i} e_{h-i} s_i$$

Claim: For all $1 \leq h \leq k$, $s_h \in \mathbb{Z}$.

Proof. We apply strong induction on h.

Base Case:

$$h = 1 \implies s_h = s_1 = (-1)^{1-1}(1)e_1 + \sum_{i=1}^{1-1}(-1)^{1-1+i}e_{1-i}s_i = (-1)^0(1)e_1 = e_1 \in \mathbb{Z}$$

since $e_i \in \mathbb{Z}$ for all $1 \leq i \leq k$, so the claim holds for the base case of $h = 1$.

Inductive Hypothesis: Assume $s_h \in \mathbb{Z}$ for all $1 \leq h \leq m < k$.

Inductive Step: Consider $h = m + 1$.

$$s_h = s_{m+1} = (-1)^{m+1-1}(m+1)e_{m+1} + \sum_{i=1}^{m+1-1}(-1)^{m+1-1+i}e_{m+1-i}s_i = (-1)^m(m+1)e_{m+1} + \sum_{i=1}^{m}(-1)^{m+i}e_{m+1-i}s_i$$

Since $(-1)^m, (m+1), e_{m+1} \in \mathbb{Z}$, we know $(-1)^m(m+1)e_{m+1} \in \mathbb{Z}$.

Also, $(-1)^{m+i}, e_{m+1-i} \in \mathbb{Z}$ for all $1 \leq i \leq m$, and the Inductive Hypothesis guarantees that $s_i \in \mathbb{Z}$ for all $1 \leq i \leq m$,

$$\implies (-1)^{m+i}e_{m+1-i}s_i \in \mathbb{Z} \text{ for all } 1 \leq i \leq m \implies \sum_{i=1}^{m}(-1)^{m+i}e_{m+1-i}s_i \in \mathbb{Z}$$

Therefore, s_{m+1} is the sum of two integers, so $s_{m+1} \in \mathbb{Z}$, which is exactly what we want to show.

The conclusion follows by strong induction. □

Newton's Identities also guarantee that, for all $h > k$, $h \in \mathbb{N}$,

$$s_h := \sum_{j=1}^{k} z_j^h = \sum_{i=h-k}^{h-1}(-1)^{h-1+i}e_{h-i}s_i$$

Claim: For all $h > k$, $s_h \in \mathbb{Z}$.

Proof. Again, we apply strong induction on h.

Base Case: $h = k + 1$.

$$s_h = s_{k+1} = \sum_{i=k+1-k}^{k+1-1}(-1)^{k+1-1+i}e_{k+1-i}s_i = \sum_{i=1}^{k}(-1)^{k+i}e_{k+1-i}s_i$$

We know $(-1)^{k+i}, e_{k+1-i} \in \mathbb{Z}$ for all $1 \leq i \leq k$, and by the previous proof, we know $s_i \in \mathbb{Z}$ for all $1 \leq i \leq k$

$$\implies (-1)^{k+i}e_{k+1-i}s_i \in \mathbb{Z} \text{ for all } 1 \leq i \leq k \implies s_{k+1} = \sum_{i=1}^{k}(-1)^{k+i}e_{k+1-i}s_i \in \mathbb{Z},$$

so the claim holds for the base case of $h = k + 1$.

Inductive Hypothesis: Assume $s_h \in \mathbb{Z}$ for all $k + 1 \leq h \leq m$.

Inductive Step: Consider $h = m + 1$.

$$s_h = s_{m+1} = \sum_{i=m+1-k}^{m+1-1}(-1)^{m+1-1+i}e_{m+1-i}s_i = \sum_{i=m+1-k}^{m}(-1)^{m+i}e_{m+1-i}s_i$$

We know $(-1)^{m+i}, e_{m+1-i} \in \mathbb{Z}$ for all $m+1-k \leq i \leq m$.

Also, combining our inductive hypothesis with the previous proof, we know

$$s_i \in \mathbb{Z} \text{ for all } 1 \leq i \leq m \implies s_i \in \mathbb{Z} \text{ for all } m+1-k \leq i \leq m$$

Thus,

$$(-1)^{m+i} e_{m+1-i} s_i \in \mathbb{Z} \text{ for all } m+1-k \leq i \leq m \implies s_{m+1} = \sum_{i=m+1-k}^{m} (-1)^{m+i} e_{m+1-i} s_i \in \mathbb{Z},$$

which is exactly what we want to show. The conclusion follows by strong induction. □

Combining these two inductive proofs, we know

$$s_h := \sum_{j=1}^{k} z_j^h \in \mathbb{Z} \text{ for all } h \in \mathbb{N},$$

so we know that

$$\sum_{j=1}^{k} z_j^{\ell n-1} = (z_1^{\ell n-1} + ... + z_k^{\ell n-1}) \in \mathbb{Z} \text{ for all } \ell \geq 2, n \in \mathbb{N},$$

so we know that

$$n | \sum_{d|n} \sum_{j=1}^{k} z_j^{\ell-1} p_d'(z_j^\ell) \text{ for all } \ell \geq 2, n \in \mathbb{N},$$

which concludes Case 1.

Case 2: $\ell = 1$. Note that

$$z_j^{\ell-1} p_d'(z_j^\ell) = \frac{\frac{d}{dz_j}(p_d(z_j^\ell))}{\ell} = \frac{d}{dz_j}(p_d(z_j)) = p_d'(z_j) = \sum_{q|d} \mu(\frac{d}{q}) q (z_j)^{q-1} = \sum_{q|d} \mu(\frac{d}{q}) q (z_j^\ell)^{q-1}$$

Therefore, the right side of our divisibility statement becomes

$$\sum_{d|n} \sum_{j=1}^{k} z_j^{\ell-1} p_d'(z_j^\ell) = \sum_{j=1}^{k} \sum_{d|n} \sum_{q|d} \mu(\frac{d}{q}) q (z_j^\ell)^{q-1} = \sum_{j=1}^{k} \sum_{d|n} \sum_{q|d} \mu(\frac{d}{q}) p_{z_j}(q)$$

Applying the Möbius Inversion Formula exactly the same way as in Case 1, we find that

$$\sum_{j=1}^{k} \sum_{d|n} \sum_{q|d} \mu(\frac{d}{q}) p_{z_j}(q) = \sum_{j=1}^{k} \sum_{d|n} f_{z_j}(d) = \sum_{j=1}^{k} p_{z_j}(n) = \sum_{j=1}^{k} n (z_j^\ell)^{n-1} = n \sum_{j=1}^{k} z_j^{n-1}$$

Note that

$$n | \sum_{d|n} \sum_{j=1}^{k} z_j^{\ell-1} p_d'(z_j^\ell) = n \sum_{j=1}^{k} z_j^{n-1} \iff \sum_{j=1}^{k} z_j^{n-1} = (z_1^{n-1} + ... + z_k^{n-1}) \in \mathbb{Z},$$

for all $n \in \mathbb{N}$. so it suffices to show that

$$\sum_{j=1}^{k} z_j^{n-1} = (z_1^{n-1} + ... + z_k^{n-1}) \in \mathbb{Z}$$

for all $n \in \mathbb{N}$.

Case 2(a): $n \geq 2$. Since $n \geq 2$, we know that $n-1 \geq 1 \implies n-1 \in \mathbb{N}$, so it suffices to show

$$\sum_{j=1}^{k} z_j^h \in \mathbb{Z} \text{ for all } h \in \mathbb{N},$$

which we already did in Case 1, so we know

$$n | \sum_{d|n} \sum_{j=1}^{k} z_j^{\ell-1} p_d'(z_j^\ell) = n \sum_{j=1}^{k} z_j^{n-1} \text{ for all } n \geq 2 \text{ when } \ell = 1,$$

which concludes Case 2(a).

Case 2(b): $n = 1$. Note that, for all $a \in \mathbb{Z}$, $n = 1 | a$, so it suffices to show

$$\sum_{d|n} \sum_{j=1}^{k} z_j^{\ell-1} p_d'(z_j^\ell) \in \mathbb{Z} \text{ when } \ell = n = 1$$

Recall that, since $\ell = 1$, we know

$$\sum_{d|n} \sum_{j=1}^{k} z_j^{\ell-1} p_d'(z_j^\ell) = \sum_{j=1}^{k} \sum_{d|n} \frac{\frac{d}{dz_j}(p_d(z_j^\ell)}{\ell} = \sum_{j=1}^{k} \sum_{d|n} \frac{d}{dz_j}(p_d(z_j)) = \sum_{j=1}^{k} \sum_{d|n} \frac{d}{dz_j}(\sum_{q|d} \mu(\frac{d}{q}) z_j^q)$$

Since $n = 1$, $d = 1$ is n's only positive divisor, and $q = 1$ is d's only positive divisor. Therefore,

$$\sum_{d|n} \sum_{j=1}^{k} z_j^{\ell-1} p_d'(z_j^\ell) = \sum_{j=1}^{k} \sum_{d|n} \frac{d}{dz_j}(\sum_{q|d} \mu(\frac{d}{q}) z_j^q) = \sum_{j=1}^{k} \frac{d}{dz_j}(\mu(\frac{1}{1}) z_j^1) = \sum_{j=1}^{k} \frac{d}{dz_j}(z_j) = \sum_{j=1}^{k} (1) = k \in \mathbb{Z},$$

so we know

$$n | \sum_{d|n} \sum_{j=1}^{k} z_j^{\ell-1} p_d'(z_j^\ell) \text{ when } \ell = n = 1,$$

which concludes Case 2(b).

Combining Case 1, Case 2(a), and Case 2(b), we have shown that

$$n | \sum_{d|n} \sum_{j=1}^{k} z_j^{\ell-1} p_d'(z_j^\ell) \text{ for all } \ell, n \in \mathbb{N}$$

which is exactly what we want to show, and thus concludes the proof. □

Assignment 11

Note: For the entirety of this assignment, $(\frac{a}{b})$ refers to the Legendre Symbol, not the fraction $\frac{a}{b}$.

Problem 1

(5 points). (a) Is 17 a quadratic residue modulo 37?

Claim: 17 is a quadratic non-residue residue modulo 37.

Proof. It suffices to show that $(\frac{17}{37}) = -1$.

Since 17 and 37 are distinct primes, the Quadratic Reciprocity of Gauss guarantees that

$$(\frac{17}{37}) = (-1)^{\frac{17-1}{2}\frac{37-1}{2}}(\frac{37}{17}) = (-1)^{8*18}(\frac{37}{17}) = (\frac{3}{17})$$

Since 3 and 17 are distinct primes, we can apply the Quadratic Reciprocity of Gauss again to find

$$(\frac{17}{37}) = (\frac{3}{17}) = (-1)^{\frac{3-1}{2}\frac{17-1}{2}}(\frac{17}{3}) = (-1)^{1*8}(\frac{17}{3}) = (\frac{2}{3})$$

Since, for all $x \in \mathbb{Z}$, $x^2 \not\equiv 2 \pmod 3$, we know

$$x^2 \equiv 2 \pmod 3$$

has no integer solutions, so we know

$$(\frac{17}{37}) = (\frac{3}{17}) = (\frac{2}{3}) = -1,$$

which is exactly what we want to show, and thus concludes the proof. □

(b) Is 35 a square modulo 41?

Claim: 35 is *not* a square modulo 41.

Proof. 35 is a square modulo 41 $\iff$ 35 is a quadratic residue modulo 41, so it suffices to show that $(\frac{35}{41}) = -1$.

Since $(\frac{ab}{d}) = (\frac{a}{d})(\frac{b}{d})$, and $35 = 5 * 7$, we know that

$$(\frac{35}{41}) = (\frac{5}{41})(\frac{7}{41})$$

Since 5, 7, and 41 are all distinct primes, the Quadratic Reciprocity of Gauss guarantees that

$$(\frac{35}{41}) = (\frac{5}{41})(\frac{7}{41}) = (-1)^{\frac{5-1}{2}\frac{41-1}{2}}(\frac{41}{5})(-1)^{\frac{7-1}{2}\frac{41-1}{2}}(\frac{41}{7}) = (-1)^{2*20+3*20}(\frac{41}{5})(\frac{41}{7}) = (\frac{1}{5})(\frac{-1}{7})$$

$(\frac{1}{p}) = 1$ for all odd primes p, so we know $(\frac{1}{5}) = 1$,

$$\implies (\frac{35}{41}) = (\frac{1}{5})(\frac{-1}{7}) = (\frac{-1}{7}) = (-1)^{\frac{7-1}{2}} = (-1)^3 = -1,$$

which is exactly what we want to show, and thus concludes the proof. □

Problem 2

(5 points). Show that if $p = x^2 - 2y^2$, $(x, y \in \mathbb{Z})$ is a prime, then $p = 2$ or $p \equiv \pm 1 \pmod 8$.

Proof. Note: let $x = 2$, $y = 1$, and we can see that $x^2 - 2y^2 = 2^2 - 2(1)^2 = 4 - 2 = 2 = p$, so,

if $p = x^2 - 2y^2$, $(x, y \in \mathbb{Z})$ is a prime, then $p = 2$ is a possibility.

Thus, it suffices to show that, if $p = x^2 - 2y^2$, $(x, y \in \mathbb{Z})$ is an odd prime, then $p \equiv \pm 1 \pmod 8$.

Assume $p = x^2 - 2y^2$, $(x, y \in \mathbb{Z})$ is an odd prime.

$$\implies x^2 - 2y^2 \equiv 0 \text{ (mod p)} \implies x^2 \equiv 2y^2 \text{ (mod p)} \tag{29}$$

If $p|y$, then $x^2 \equiv 0 \pmod{p} \implies p|x^2 \implies p|x$.

Therefore, $p^2|x^2$ and $p^2|y^2 \implies p^2|x^2 - 2y^2 = p \implies p \equiv 0 \pmod{p^2}$, which is a contradiction.

Therefore, $p \nmid y$, so y is invertible modulo p, so y^{-1} exists modulo p.

Multiplying both sides of (1) by $(y^{-1})^2$, we find

$$(xy^{-1})^2 \equiv 2 \pmod{p} \implies (\frac{2}{p}) = 1$$

since for all odd primes p, $p \nmid 2$. Also, since p is an odd prime, we know

$$(\frac{2}{p}) = 1 = (-1)^{\frac{p^2-1}{8}} \implies \frac{p^2-1}{8} = 2k \implies p^2 - 1 = 16k \text{ for some } k \in \mathbb{Z}$$

Therefore,

$$16|p^2 - 1 = (p+1)(p-1) \implies (p-1)(p+1) \equiv 0 \pmod{16},$$

and we want to show that $p \equiv \pm 1 \pmod{8}$.

To do so, assume to the contrary that $p \not\equiv \pm 1 \pmod{8}$. Since p is an odd prime, we know $p \equiv \pm 3 \pmod{8}$. Therefore,

$$(p+1)(p-1) \equiv (-1)^2(4)(2) = 8 \pmod{16}, \text{ which is a contradiction.}$$

Thus, if p is an odd prime such that $p = x^2 - 2y^2$ for some $x, y \in \mathbb{Z}$, then $p \equiv \pm 1 \pmod{8}$.

Thus, if $p = x^2 - 2y^2$, $(x, y \in \mathbb{Z})$ is a prime, then $p = 2$ or $p \equiv \pm 1 \pmod{8}$, which is exactly what we want to show,

and thus concludes the proof. □

Problem 3

(5 points). Show that if $p = x^2 + 3y^2$ $(x, y \in \mathbb{Z})$ is a prime, then $p = 3$ or $p \equiv 1 \pmod{3}$.

Proof. First, we must show that $2 = x^2 + 3y^2$ has no integer solutions $x, y \in \mathbb{Z}$. This is trivially true if $x = y = 0$.

If $x = 0, y \neq 0$, we know that $2 = x^2 + 3y^2 = 3y^2 \implies \frac{2}{3} = y^2$, which has no integer solutions $y \in \mathbb{Z}$.

If $x \neq 0, y = 0$, we know that $2 = x^2 + 3y^2 = x^2$, which has no integer solutions $x \in \mathbb{Z}$.

Finally, if $x, y \neq 0$, we know that, if $x, y \in \mathbb{Z}$, $2 = x^2 + 3y^2 \geq 4$, which is a contradiction.

Thus, we have shown that $2 = x^2 + 3y^2$ has no integer solutions, so we now just have to consider odd primes.

Also, if we let $x = 0$, $y = 1$, then $x^2 + 3y^2 = 3 = p$, so,

if $p = x^2 + 3y^2$ $(x, y \in \mathbb{Z})$ is a prime, then $p = 3$ is a possibility.

Now, it suffices to show that, if $p = x^2 + 3y^2 \neq 3$ $(x, y \in \mathbb{Z})$ is an odd prime, then $p \equiv 1 \pmod{3}$.

Assume $p \neq 3$ is an odd prime such that $p = x^2 + 3y^2$.

$$\implies x^2 + 3y^2 \equiv 0 \pmod{p} \implies x^2 \equiv -3y^2 \pmod{p} \tag{30}$$

If $p|y$, then $x^2 \equiv 0 \pmod{p} \implies p|x^2 \implies p|x$, so we know

$$p^2|x^2, p^2|y^2 \implies p^2|x^2+3y^2 = p \implies p \equiv 0 \pmod{p^2}, \text{ which is a contradiction.}$$

Thus $p \nmid y \implies$ y is invertible modulo p, so y^{-1} exists modulo p. Multiplying both sides of (2) by $(y^{-1})^2$, we find

$$(xy^{-1})^2 \equiv -3 \pmod{p} \implies (\frac{-3}{p}) = 1$$

Since $(\frac{ab}{d}) = (\frac{a}{d})(\frac{b}{d})$, and $-3 = (-1)(3)$, we know

$$(\frac{-3}{p}) = (\frac{-1}{p})(\frac{3}{p}).$$

Since $p \neq 3$, we know p and 3 are distinct odd primes, so the Quadratic Reciprocity of Gauss guarantees that

$$(\frac{-3}{p}) = (\frac{-1}{p})(\frac{3}{p}) = (-1)^{\frac{p-1}{2}}(-1)^{\frac{3-1}{2}\frac{p-1}{2}}(\frac{p}{3}) = (-1)^{\frac{p-1}{2}+\frac{p-1}{2}}(\frac{p}{3}) = (-1)^{p-1}(\frac{p}{3}) = (\frac{p}{3})$$

since p being odd guarantees that $p-1$ is even. $\implies x^2 \equiv p \pmod{3}$ has an integer solution $\implies p \equiv 1 \pmod{3}$,

since $p \neq 3 \implies p \neq 0 \pmod{3}$, and $x^2 \equiv 0$ or $1 \pmod{3}$ for all $x \in \mathbb{Z}$.

Thus, we have shown that, if $p = x^2 + 3y^2 \neq 3$ $(x, y \in \mathbb{Z})$ is an odd prime, then $p \equiv 1 \pmod{3}$,

which is exactly what we want to show, and thus concludes the proof. □

Problem 4

(5 points). Show that if $p = x^2 + xy + 3y^2$, $(x, y \in \mathbb{Z})$ is a prime, then $p = 11$ or $p \equiv 1$, 3, 4, 5, or 9 (mod 11).

Proof. First, we need to show that $2 = x^2 + xy + 3y^2$ has no integer solutions $x, y \in \mathbb{Z}$.

Case 1: x, y are both even, so we can write $x = 2k$ and $y = 2s$ for some $k, s \in \mathbb{Z}$.

$$\implies x^2 + xy + 3y^2 = 4k^2 + 4ks + 12s^2 = 4(k^2 + ks + 3s^2) \neq 2$$

since $k^2 + ks + 3s^2 \in \mathbb{Z}$ and $4q \neq 2$ for all $q \in \mathbb{Z}$.

Case 2: x, y are both odd, so we can write $x = 2k+1$ and $y = 2s+1$ for some $k, s \in \mathbb{Z}$.

$$\begin{aligned} \implies x^2 + xy + 3y^2 = &\quad (2k+1)^2 + (2k+1)(2s+1) + 3(2s+1)^2 = 4k^2 + 4k + 1 + 4ks + 2k + 2s + 1 + 12s^2 + 12s + 3 \\ = &\quad 4k^2 + 6k + 12s^2 + 14s + 5 = 2(2k^2 + 3k + 6s^2 + 7s + 2) + 1 \neq 2 \end{aligned}$$

since 2 is even but $2(2k^2 + 3k + 6s^2 + 7s + 2) + 1$ is odd for all $k, s \in \mathbb{Z}$.

Case 3: x is odd, y is even, so we can write $x = 2k+1$ and $y = 2s$ for some $k, s \in \mathbb{Z}$.

$\implies x^2+xy+3y^2 = (2k+1)^2+(2k+1)(2s)+3(2s)^2 = 4k^2+4k+1+4ks+2s+12s^2 = 2(2k^2+2k+2ks+s+6s^2)+1 \neq 2$

since 2 is even but $2(2k^2 + 2k + 2ks + s + 6s^2) + 1$ is odd for all $k, s \in \mathbb{Z}$.

Case 4: x is even, y is odd, so we can write $x = 2k$ and $y = 2s+1$ for some $k, s \in \mathbb{Z}$.

$\implies x^2+xy+3y^2 = (2k)^2+(2k)(2s+1)+3(2s+1)^2 = 4k^2+4ks+2k+12s^2+12s+3 = 2(2k^2+2ks+k+6s^2+6s+1)+1 \neq 2$

since 2 is even but $2(2k^2+2ks+k+6s^2+6s+1)+1$ is odd for all $k, s \in \mathbb{Z}$.

Thus, we have shown that if $p = x^2+xy+3y^2$, $(x, y \in \mathbb{Z})$ is a prime, then $p \neq 2$.

Note: If we let $x = -1$ and $y = 2$, then $x^2+xy+3y^2 = (-1)^2+(-1)(2)+3(2)^2 = 1-2+12 = 11$, so

if $p = x^2+xy+3y^2$ $(x, y \in \mathbb{Z})$ is a prime, then $p = 11$ is a possibility.

Now, it suffices to show that, if $p = x^2+xy+3y^2 \neq 11$ $(x, y \in \mathbb{Z})$ is an odd prime, then $p \equiv 1$, 3, 4, 5, or 9 (mod 11).

Assume $p = x^2+xy+3y^2 \neq 11$ is an odd prime.

$$\implies 4p = 4x^2+4xy+12y^2 = (2x)^2+2(2x)y+y^2+11y^2 = (2x+y)^2+11y^2 \equiv 0 \text{ (mod p)}$$

$$\implies (2x+y)^2 \equiv -11y^2 \text{ (mod p)} \tag{31}$$

If $p|y$, then $y \equiv 0$ (mod p),

$$\implies (2x+y)^2 \equiv 4x^2 \equiv 0 \text{ (mod p)} \implies p|x^2 \implies p|x$$

Thus, $p^2|x^2$, $p^2|y^2$, and $p^2|xy$, so

$$p^2|x^2+xy+3y^2 = p \implies p \equiv 0 \text{ (mod } p^2),$$

which is a contradiction.

Therefore, $p \nmid y$, so y is invertible modulo p, so y^{-1} exists modulo p.

Multiplying both sides of (3) by $(y^{-1})^2$, we obtain

$$((2x+y)y^{-1})^2 \equiv -11 \text{ mod p} \implies (\frac{-11}{p}) = 1$$

since $p \neq 11$.Since $(\frac{ab}{d}) = (\frac{a}{d})(\frac{b}{d})$, and $-11 = -1*11$, we know that

$$(\frac{-11}{p}) = 1 = (\frac{-1}{p})(\frac{11}{p})$$

Since $p \neq 11$, 11 and p are distinct odd primes, so the Quadratic Reciprocity of Gauss guarantees that

$$(\frac{-11}{p}) = 1 = (\frac{-1}{p})(\frac{11}{p}) = (-1)^{\frac{p-1}{2}}(-1)^{\frac{11-1}{2}\frac{p-1}{2}}(\frac{p}{11}) = (-1)^{\frac{6(p-1)}{2}}(\frac{p}{11}) = (-1)^{3(p-1)}(\frac{p}{11}) = (\frac{p}{11})$$

since p is odd $\implies p-1$ is even.

Thus, if $p = x^2+xy+3y^2 \neq 11$ $(x, y \in \mathbb{Z})$ is an odd prime, then p must be a nonzero quadratic residue modulo 11.

Note:

$1^2 = \quad 1 \equiv 1 \text{ (mod 11)}, 2^2 = 4 \equiv 4 \text{ (mod 11)}, 3^2 = 9 \equiv 9 \text{ (mod 11)}, 4^2 = 16 \equiv 5 \text{ (mod 11)}, 5^2 = 25 \equiv 3 \text{ (mod 11)},$
$6^2 = \quad 36 \equiv 3 \text{ (mod 11)}, 7^2 = 49 \equiv 5 \text{ (mod 11)}, 8^2 = 64 \equiv 9 \text{ (mod 11)}, 9^2 = 81 \equiv 4 \text{ (mod 11)}, 10^2 = 100 \equiv 1 \text{ (mod 11)}$

So all nonzero quadratic residues modulo 11 $\equiv$ 1, 3, 4, 5, or 9 (mod 11).

Therefore, if $p = x^2+xy+3y^2 \neq 11$ $(x, y \in \mathbb{Z})$ is an odd prime, then $p \equiv 1$, 3, 4, 5 or 9 (mod 11),

which is exactly what we want to show, and thus concludes the proof. $\square$

Problem 5

(Bonus, 5 points). Does

$$x_1^{169} + ... + x_{666}^{169} = (x_1 + ... + x_{666})^2 + 2$$

have integer solutions $(x_1, ..., x_{666}) \in \mathbb{Z}^{666}$?

Claim:

$$x_1^{169} + ... + x_{666}^{169} = (x_1 + ... + x_{666})^2 + 2$$

does not have integer solutions $(x_1, ..., x_{666}) \in \mathbb{Z}^{666}$.

Proof. Note that

$$x_1^{169} + ... + x_{666}^{169} = (x_1^{13})^{13} + ... + (x_{666}^1 3)^1 3 \equiv x_1^{13} + ... + x_{666}^{13} \equiv x_1 + ... + x_{666} \pmod{13}$$

since Fermat's little theorem guarantees that $a^p \equiv a \pmod{p}$ for all $a \in \mathbb{Z}$, p prime.

Thus, it suffices to show

$$x_1 + ... + x_{666} \not\equiv (x_1 + ... + x_{666})^2 + 2 \pmod{13}$$

for all $(x_1, ..., x_{666}) \in \mathbb{Z}^{666}$.

We can do this by running through all 13 possible options for $(x_1 + ... + x_{666}) \pmod{13}$.

Case 1:

$$x_1 + ... + x_{666} \equiv 0 \pmod{13} \implies (x_1 + ... + x_{666})^2 + 2 \equiv 0^2 + 2 = 2 \not\equiv 0 \pmod{13}$$

Case 2:

$$x_1 + ... + x_{666} \equiv 1 \pmod{13} \implies (x_1 + ... + x_{666})^2 + 2 \equiv 1^2 + 2 = 3 \not\equiv 1 \pmod{13}$$

Case 3:

$$x_1 + ... + x_{666} \equiv 2 \pmod{13} \implies (x_1 + ... + x_{666})^2 + 2 \equiv 2^2 + 2 = 6 \not\equiv 2 \pmod{13}$$

Case 4:

$$x_1 + ... + x_{666} \equiv 3 \pmod{13} \implies (x_1 + ... + x_{666})^2 + 2 \equiv 3^2 + 2 = 11 \not\equiv 3 \pmod{13}$$

Case 5:

$$x_1 + ... + x_{666} \equiv 4 \pmod{13} \implies (x_1 + ... + x_{666})^2 + 2 \equiv 4^2 + 2 = 18 \equiv 5 \not\equiv 4 \pmod{13}$$

Case 6:

$$x_1 + ... + x_{666} \equiv 5 \pmod{13} \implies (x_1 + ... + x_{666})^2 + 2 \equiv 5^2 + 2 = 27 \equiv 1 \not\equiv 5 \pmod{13}$$

Case 7:

$$x_1 + ... + x_{666} \equiv 6 \pmod{13} \implies (x_1 + ... + x_{666})^2 + 2 \equiv 6^2 + 2 = 38 \equiv 12 \not\equiv 6 \pmod{13}$$

Case 8:

$$x_1 + ... + x_{666} \equiv 7 \pmod{13} \implies (x_1 + ... + x_{666})^2 + 2 \equiv 7^2 + 2 = 51 \equiv 12 \not\equiv 7 \pmod{13}$$

Case 9:

$$x_1 + ... + x_{666} \equiv 8 \pmod{13} \implies (x_1 + ... + x_{666})^2 + 2 \equiv 8^2 + 2 = 66 \equiv 1 \not\equiv 8 \pmod{13}$$

Case 10:

$$x_1 + ... + x_{666} \equiv 9 \pmod{13} \implies (x_1 + ... + x_{666})^2 + 2 \equiv 9^2 + 2 = 83 \equiv 5 \not\equiv 9 \pmod{13}$$

Case 11:

$$x_1 + ... + x_{666} \equiv 10 \pmod{13} \implies (x_1 + ... + x_{666})^2 + 2 \equiv 10^2 + 2 = 102 \equiv 11 \not\equiv 10 \pmod{13}$$

Case 12:

$$x_1 + ... + x_{666} \equiv 11 \pmod{13} \implies (x_1 + ... + x_{666})^2 + 2 \equiv 11^2 + 2 = 123 \equiv 6 \not\equiv 11 \pmod{13}$$

Case 13:

$$x_1 + ... + x_{666} \equiv 12 \pmod{13} \implies (x_1 + ... + x_{666})^2 + 2 \equiv 12^2 + 2 = 146 \equiv 3 \not\equiv 12 \pmod{13}$$

Thus, we have shown that

$$x_1 + ... + x_{666} \not\equiv (x_1 + ... + x_{666})^2 + 2 \pmod{13}$$

for all $(x_1, ..., x_{666}) \in \mathbb{Z}^{666}$, which is exactly what we want to show,

and thus concludes the proof. $\square$