CS70 Discrete Mathematics and Probability Theory, Fall 2018

Final Exam

8:00-11:00am, 14 December

Your First Name: **Your Last Name:**

SIGN Your Name: **Your SID Number:**

Your Exam Room:

Name of Person Sitting on Your Left:

Name of Person Sitting on Your Right:

Name of Person Sitting in Front of You:

Name of Person Sitting Behind You:

Instructions:

(a) *As soon as the exam starts, please write your student ID in the space provided at the top of every page! (We will remove the staple when scanning your exam.)*

(b) *There are 9* **double-sided** *sheets (18 numbered pages) on the exam. Notify a proctor immediately if a sheet is missing.*

(c) *We will not grade anything outside of the space provided for a question (i.e., either a designated box if it is provided, or otherwise the white space immediately below the question).* **Be sure to write your full answer in the box or space provided!** *Scratch paper is provided on request; however, please bear in mind that nothing you write on scratch paper will be graded!*

(d) *The questions vary in difficulty, so if you get stuck on any question it may help to leave it and return to it later.*

(e) *On questions 1-2: You must give the answer in the format requested (e.g., True/False, an expression, a statement.) An expression may simply be a number or an expression with a relevant variable in it. For short answer questions, correct, clearly identified answers will receive full credit with no justification. Incorrect answers may receive partial credit.*

(f) *On questions 3-8, you should give arguments, proofs or clear descriptions if requested. If there is a box you must use it for your answer.*

(g) *You may consult three two-sided "cheat sheets" of notes. Apart from that, you may not look at any other materials. Calculators, phones, computers, and other electronic devices are NOT permitted.*

(h) *You may, without proof, use theorems and lemmas that were proven in the notes and/or in lecture.*

(i) *You have 3 hours: there are 8 questions on this exam worth a total of 190 points.*

[Exam starts on next page]

1. True/False [*No justification; answer by shading the correct bubble. Points per answer as indicated; total of 39 points. No penalty for incorrect answers.*]

(a) Zero or more of the following are valid logical equivalences, for arbitrary propositions P and Q. Indicate which by shading the appropriate circles.

YES	NO		
●	○	$(P \Rightarrow \neg Q) \equiv (Q \Rightarrow \neg P)$	*1pt*
○	●	$(P \Rightarrow Q) \equiv (P \vee \neg Q)$	*1pt*
●	○	$\neg(\neg P \vee Q) \equiv (P \wedge \neg Q)$	*1pt*
○	●	$((P \Rightarrow Q) \wedge (Q \Rightarrow P)) \equiv ((P \wedge \neg Q) \vee (\neg P \wedge Q))$	*1pt*
●	○	$\neg(P \vee (\neg P \wedge Q)) \equiv (\neg P \wedge \neg Q)$	*1pt*

(b) Define the following predicates involving the variables x, y over the universe of cats.

- $G(x)$: x has green eyes
- $B(x)$: x has a bushy tail
- $F(x, y)$: x is fatter than y

Consider the following statement:

"Every cat with green eyes is fatter than at least one cat that has a bushy tail but doesn't have green eyes."

Which (if any) of the following expressions are accurate translations of this statement? Answer by shading either the "Yes" or the "No" bubble for each expression. (There may be more than one "Yes" answer.)

YES	NO		
○	●	$\forall x \exists y (G(x) \wedge B(y) \wedge \neg G(y) \wedge F(x, y))$	*1pt*
○	●	$\exists y \forall x (\neg G(y) \wedge B(y) \wedge F(y, x))$	*1pt*
○	●	$\forall x \forall y (G(x) \Rightarrow (B(y) \wedge \neg G(y) \wedge F(x, y)))$	*1pt*
●	○	$\forall x \exists y (G(x) \Rightarrow (B(y) \wedge \neg G(y) \wedge F(x, y)))$	*1pt*

[Q1 continued on next page]

(c) Consider the following stable marriage instance, consisting of four men 1, 2, 3, 4 and four women A, B, C, D:

Man	Women			
1	A	B	D	C
2	D	C	B	A
3	A	C	D	B
4	B	A	C	D

Woman	Men			
A	2	3	1	4
B	2	3	1	4
C	1	4	2	3
D	4	2	1	3

For each of the following statements about this instance, indicate whether the statement is True or False by shading the corresponding bubble.

TRUE FALSE

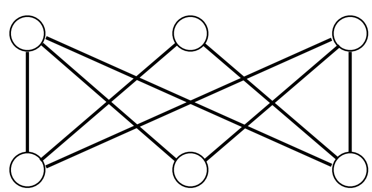
○ ● The pairing (1,D), (2, C), (3, A), (4, B) is stable. *2pts*

● ○ The pairing (1, C), (2, B), (3, A), (4, D) is male pessimal. *2pts*

○ ● There exists a stable pairing in which man 4 is paired with woman A. *2pts*

● ○ Woman A is paired with man 3 in every stable pairing. *2pts*

(d) Consider the following graph.



Which of the following properties is/are true of this graph?

TRUE FALSE

● ○ The graph is connected. *1pt*

● ○ The graph is bipartite. *1pt*

● ○ The graph is planar. *1pt*

● ○ The graph has a Hamiltonian cycle. *1pt*

[Q1 continued on next page]

(e) A connected graph G (with no self-loops or multiple edges between the same pair of vertices) has seven vertices whose degrees are 3, 4, 4, 4, 5, 6, 6 respectively. Answer each of the following by shading the appropriate bubble: "Yes", "No" or "??" if the answer cannot be determined from the given information.

YES	**NO**	**??**		
○	●	○	Does G have an Eulerian tour?	*1pt*
●	○	○	Does G have an Eulerian walk?	*1pt*
○	●	○	Is G planar?	*1pt*

(f) Classify each of the following functions $f : \mathbb{R} \to \mathbb{R}$ as (i) neither 1-1 nor onto; (ii) 1-1 but not onto; (iii) onto but not 1-1; (iv) both 1-1 and onto (a bijection).

(i)	**(ii)**	**(iii)**	**(iv)**		
●	○	○	○	$f(x) = x^2$.	*1pt*
○	○	○	●	$f(x) = x + 1$.	*1pt*
○	○	○	●	$f(x) = \frac{1}{x}$ for $x \neq 0$, $f(0) = 0$.	*1pt*
○	●	○	○	$f(x) = e^x$.	*1pt*

(g) Answer each of the following questions **TRUE** or **FALSE** by shading the appropriate bubble.

TRUE	**FALSE**		
●	○	There are two distinct powers of 2 that are equal mod 97.	*1pt*
●	○	For sets A, B, if A is uncountable and B is countable, then the difference $A \setminus B$ is uncountable.	*1pt*
○	●	For a collection of countably infinite sets A_i, $i \in \mathbb{N}$, the union $\bigcup_{i \in \mathbb{N}} A_i$ is uncountable.	*1pt*
○	●	If A is uncountable and B is finite and non-empty, then $A \times B$ is countable.	*1pt*

[Q1 continued on next page]

(h) Answer each of the following questions **TRUE** or **FALSE** by shading the appropriate bubble.

TRUE	FALSE		
●	○	For all events $A_1, \ldots, A_n$, we have $\mathbb{P}[\bigcup_{i=1}^n A_i] \leq \sum_{i=1}^n \mathbb{P}[A_i]$.	*1pt*
○	●	For dependent random variables X, Y and constants a, b, it is possible that $\mathbb{E}[aX + bY] \neq a\mathbb{E}[X] + b\mathbb{E}[Y]$.	*1pt*
○	●	$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if and only if X and Y are independent.	*1pt*
○	●	Consider two random variables X and Y with ranges $\mathcal{A}_X$ and $\mathcal{A}_Y$, respectively. If there exist $a \in \mathcal{A}_X$ and $b \in \mathcal{A}_Y$ such that $\mathbb{P}[X = a, Y = b] = \mathbb{P}[X = a]\mathbb{P}[Y = b]$, then X and Y are independent.	*1pt*
●	○	Consider the standard Coupon Collector's Problem with n coupon types, and let W_n denote the total number of trials required to collect all n coupon types. Then, $\lim_{n\to\infty} \mathbb{E}[W_n]/(n \ln n) = 1$.	*1pt*
●	○	Suppose $P(x) = Ax + B$ is a random polynomial where A and B are independent standard normal random variables. Then $P(x)$ has a root with probability 1.	*1pt*
●	○	For all Markov chains $\{X_n, n \in \mathbb{N}\}$ with finite state space S and $\forall m \in \{0, 1, \ldots, n\}$, $\mathbb{P}[X_n = j \mid X_0 = i] = \sum_{k \in S} \mathbb{P}[X_n = j \mid X_m = k] \times \mathbb{P}[X_m = k \mid X_0 = i]$.	*1pt*

2. Short Answers [*Answer is a single number or expression;* **write it in the box provided: anything outside the box will not be graded;** *no justification necessary. 3 points per answer; total of 78 points. No penalty for incorrect answers.*]

(a) What is the inverse of the function $f(x) = 3x - 2$ over the reals? *3pts*

$f^{-1}(y) = (y+2)/3$. [Every $x \in \mathbb{R}$ gets mapped to a unique $y = 3x - 2$ by f, and so solving for x tells us that every $y \in \mathbb{R}$ gets uniquely mapped to $f^{-1}(y) = (y+2)/3$.]

(b) Compute $\gcd(323, 152)$. *3pts*

19. [$\gcd(323, 152) = \gcd(152, 19) = \gcd(19, 0) = 19$.]

(c) What is $37^{225} \bmod 113$? [Note: 113 is prime.] *3pts*

37. [$37^{225} \equiv 37^{2\times 112} \cdot 37 \equiv 37 \pmod{113}$ by Fermat's Little Theorem.]

(d) Solve for x in the modular equation $5x - 7 = 0 \pmod{11}$. *3pts*

$x \equiv -3 \equiv 8 \pmod{11}$. [$1 = 11 + (-2)\cdot 5$, and so $5^{-1} \equiv -2 \pmod{11}$. Multiplying the entire equation by -2 gives $x + 14 \equiv 0 \pmod{11}$.]

(e) Suppose Alice's public RSA key is $(N, e) = (143, 7)$. What is the value d that Alice needs in order to decrypt messages sent to her? (Hint: Note that $143 = 11 \times 13$.) *3pts*

$d = 103$. [$d \equiv e^{-1} \pmod{(p-1)(q-1)}$, and here $p = 11, q = 13$, so $(p-1)(q-1) = 120$. Since $1 = 120 - 17 \cdot 7$, we have $7^{-1} \equiv -17 \equiv 103 \pmod{120}$.]

(f) Rex chooses a random polynomial P of degree at most k over $\mathrm{GF}(q)$ (for a prime $q > k$), by selecting $k+1$ coefficients independently and uniformly at random from $\{0, 1, \ldots, q-1\}$.

(i) What is the probability that Rex's polynomial goes through the point $(0,0)$, i.e., that $P(0) = 0$? *3pts*

$1/q$. [Each of the q^{k+1} polynomials of degree at most k is sampled with equal probability. Since fixing the value 0 at $x = 0$ leaves k remaining points to determine a polynomial, there are q^k polynomials with $P(0) = 0$. The probability is the ratio q^k/q^{k+1}.]

(ii) What is the probability that Rex's polynomial has exactly k distinct roots? *3pts*

$\binom{q}{k}(q-1)/q^{k+1}$. [Since the total number of polynomials is q^{k+1}, we just need to show that the number of polynomials with exactly k distinct roots is $\binom{q}{k}(q-1)$. There are $\binom{q}{k}$ possibilities for choosing the roots, which determines the polynomial up to a constant factor; the constant factor can be chosen in $q-1$ ways (since it cannot be 0). Alternatively, after choosing the zeros, we can choose the value of the polynomial at one additional point, which can be done in $q-1$ ways. (Again 0 is not allowed.)]

[Q2 continued on next page]

(g) Alice wants to share a secret (a number mod 7) among her four loyal companions in such a way that any three of the companions can recover the secret, but no two of them can. All of them have taken CS70 and agree to do polynomial secret sharing with the secret stored at $P(0)$ for a suitable polynomial P over $\mathrm{GF}(7)$. Alice gives each of her companions one of the following points: $(1,1),(2,0),(3,0),(4,1)$. What is the secret? *3pts*

3. [Alice's polynomial P must have been of degree at most 2, and so we can use the first three values to find the secret $P(0)$ through Lagrange Interpolation: $P(0) = \sum_{k=1}^{3} y_k \Delta_k(0) = \Delta_1(0)$. But $\Delta_1(x) \equiv (x-2)(x-3)2^{-1} \equiv 4(x-2)(x-3) \pmod 7$, and so $P(0) \equiv 6 \cdot 4 \equiv 3 \pmod 7$.]

(h) Laurel wants to send a message consisting of 100 packets to Hardy over a noisy channel. Laurel knows that up to $\frac{1}{6}$ of the packets sent may be corrupted by the channel. Assuming Laurel uses the Berlekamp-Welch encoding scheme, how many packets must he send to ensure that Hardy will be able to recover the original message? *3pts*

150. [We need to ensure that we have twice as many redundant packets as we have errors, i.e., the number of packets m must satisfy $m - 100 \geq 2 \cdot (m/6)$, or equivalently $m \geq 150$.]

(i) We say that program P_1 *dominates* program P_2 if P_1 halts on every input on which P_2 halts. The problem *Dominates* takes as input two programs, P_1 and P_2, and decides whether P_1 dominates P_2. The following pseudo-code gives a reduction from the Halting Problem, *Halt*, to *Dominates*. Fill in the blank to make the reduction behave correctly. *3pts*

```
Test-Halt(P,x)
    let P1 be a program that, on every input, runs P on x
    let P2 be a program that, on every input, halts
    if Test-Dominates(P1,P2) then return "yes" else return "no"
```

(j) Suppose there are k keys $\{w_1, \ldots, w_k\}$ and a hash table of size n. How many distinct hash functions are there such that keys w_1 and w_2 do not get mapped to the same location of the hash table? *3pts*

$n^{k-1}(n-1)$. [There are n choices for w_1, $n-1$ choices for w_2, and n^{k-2} choices to map the remaining $k-2$ keys.]

(k) A 5-card poker hand is called a *straight* if its cards can be re-arranged to form a contiguous sequence, regardless of their suits, i.e., if the hand is of the form $\{A,2,3,4,5\}, \{2,3,4,5,6\}, \ldots,$ or $\{10, J, Q, K, A\}$. How many straight hands are there consisting of 3 black and 2 red cards? *3pts*

$10 \cdot \binom{5}{3} \cdot 2^5$. [There are 10 distinct sets of numeric values that can form a straight. Given such a set of five numbers, there are $\binom{5}{3}$ ways of choosing which ones are red and which ones are black, and given the color of a card, there are two different suits that share this colour.]

[Q2 continued on next page]

(l) Consider the complete graph K_n with $n \geq 3$ vertices, and suppose that each edge is colored blue with probability p and red with probability $1-p$, independently of all other edges. Let Δ_n denote the number of completely blue triangles resulting from this random coloring.

(i) What is $\mathbb{E}[\Delta_n]$? *3pts*

$\binom{n}{3}p^3$. [There are $\binom{n}{3}$ triangles in K_n. We can define an indicator variable I_T for each triangle T denoting the event that its edges are all colored blue. For each triangle T we have $\mathbb{P}[I_T = 1] = p^3$ so by linearity of expectation we have $\mathbb{E}[\Delta_n] = \binom{n}{3}p^3$.]

(ii) Use the union bound to find a lower bound on $\mathbb{P}[\Delta_n = 0]$. *3pts*

$1 - \binom{n}{3}p^3$. [Equivalently, we can find an upper bound on $\mathbb{P}[\Delta_n > 0] = \mathbb{P}[\bigcup_T \{I_T = 1\}]$ where the union is over all triangles T. Applying the union bound gives us the upper bound $\mathbb{P}[\bigcup_T \{I_T = 1\}] \leq \sum_T \mathbb{P}[I_T = 1] = \binom{n}{3}p^3$. Thus $\mathbb{P}[\Delta_n = 0] = 1 - \mathbb{P}[\Delta_n > 0] \geq 1 - \binom{n}{3}p^3$.]

(m) Let A and B denote two events such that $A \subset B$. Suppose $\mathbb{P}[A] = a$ and $\mathbb{P}[B] = b$, and let I_A and I_B denote the indicator random variables for A and B, respectively. Find $\mathrm{Cov}(I_A, I_B)$. *3pts*

$a(1-b)$. [We have $\mathrm{Cov}(I_A, I_B) = \mathbb{E}[I_A I_B] - \mathbb{E}[I_A]\mathbb{E}[I_B]$. Observe that since $A \subset B$, then $I_A I_B = 1$ exactly when $I_A = 1$. Thus $\mathbb{E}[I_A I_B] = \mathbb{E}[I_A] = \mathbb{P}[A] = a$. Since $\mathbb{E}[I_B] = \mathbb{P}[B] = b$, then $\mathrm{Cov}(I_A, I_B) = a - ab$.]

(n) Consider an urn with 3 blue balls and 1 red ball, and suppose you sample one ball at a time with replacement. Let X be the number of draws required until both of the colors, blue and red, have been observed at least once.

(i) What is $\mathbb{P}[X = n \mid \text{The first ball drawn is red}]$, for $n \geq 2$? *3pts*

$(\frac{1}{4})^{n-2}\frac{3}{4}$. [Given that we first drew a red ball, if we take n draws to observe both colors, then draws 2 to $n-1$ must also have been red and draw n must be blue. This occurs with probability $(\frac{1}{4})^{n-2}\frac{3}{4}$. (Observe that we are following a geometric distribution)]

(ii) What is $\mathbb{P}[X = n]$, for $n \geq 2$? *3pts*

$\frac{3+3^{n-1}}{4^n}$. [Let R denote the event that the first ball drawn is red and B denote the event that the first ball drawn is blue. Using a similar argument as in (i) we have $\mathbb{P}[X = n \mid B] = (\frac{3}{4})^{n-2}\frac{1}{4}$. Thus

$$\mathbb{P}[X = n] = \mathbb{P}[X = n \mid R] \cdot \mathbb{P}[R] + \mathbb{P}[X = n \mid B] \cdot \mathbb{P}[B] \quad (1)$$

$$= \left(\frac{1}{4}\right)^{n-1}\frac{3}{4} + \left(\frac{3}{4}\right)^{n-1}\frac{1}{4} = \frac{3+3^{n-1}}{4^n}. \quad] \quad (2)$$

[Q2 continued on next page]

(o) Suppose $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$ are independent random variables, where $\lambda, \mu > 0$. What is $\mathbb{P}[\min\{X, Y\} \leq t]$, where t is a positive constant? *3pts*

$1 - e^{-t(\lambda+\mu)}$. [$\mathbb{P}[\min\{X,Y\} \leq t] = 1 - \mathbb{P}[\min\{X,Y\} > t] = 1 - \mathbb{P}[X > t]\mathbb{P}[Y > t]$ because X and Y are independent. Thus our answer is $1 - e^{-t\lambda}e^{-t\mu} = 1 - e^{-t(\lambda+\mu)}$]

(p) Suppose $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$ are independent random variables, where $\lambda, \mu > 0$. What is the conditional probability $\mathbb{P}[X < Y \mid \min\{X, Y\} > t]$, where t is a positive constant? *3pts*

$\frac{\lambda}{\lambda+\mu}$. [By the memoryless property of exponential random variables, $\mathbb{P}[X < Y \mid \min\{X,Y\} > t] = \mathbb{P}[X < Y]$, and from Problem 3 of Discussion 12A, we know $\mathbb{P}[X < Y] = \frac{\lambda}{\lambda+\mu}$.]

(q) Suppose $X \sim \text{Normal}(1, 2)$ and $Y \sim \text{Normal}(2, 1)$ are independent random variables. What is the distribution of $2X - Y + 1$? State its name and specify its parameter(s). *3pts*

$\text{Normal}(1, 9)$. [Linear combinations of Normal random variables and constants are also distributed normally. Let $Z = 2X - Y + 1$. Then $\mathbb{E}[Z] = 2\mathbb{E}[X] - \mathbb{E}[Y] + 1 = 2 - 2 + 1 = 1$. Also, $\text{Var}[Z] = \text{Var}[2X - Y] = \text{Var}[2X] + \text{Var}[Y] = 4\text{Var}[X] + 1 = 9$.]

(r) Suppose A and B are independent $\text{Normal}(1, 1)$ random variables. Find $\mathbb{P}[2A + B \geq 4]$ in terms of the cumulative distribution function (c.d.f.) Φ of the standard normal distribution. *3pts*

$1 - \Phi\left(\frac{1}{\sqrt{5}}\right)$. [Let $Z = 2A + B$. Then $\mathbb{E}[Z] = 2\mathbb{E}[A] + \mathbb{E}[B] = 3$ and $\text{Var}[Z] = 4\text{Var}[A] + \text{Var}[B] = 5$ so Z is a $\text{Normal}(3, 5)$ random variable, and hence $\frac{Z-3}{\sqrt{5}}$ is standard normal. Then

$$\mathbb{P}[Z \geq 4] = \mathbb{P}[Z - 3 \geq 1] = \mathbb{P}\left[\frac{Z-3}{\sqrt{5}} \geq \frac{1}{\sqrt{5}}\right] = 1 - \Phi\left(\frac{1}{\sqrt{5}}\right). \quad]$$

(s) Consider randomly dropping a circular coin of radius 1 cm onto a large rectangular grid where horizontal lines are 3 cm apart, while vertical lines are 4 cm apart. What is the probability that the coin intersects at least one grid line? *3pts*

$\frac{5}{6}$. [We compute the probability that the coin intersects no grid lines. Consider the 3 by 4 rectangle between two adjacent horizontal lines and two adjacent vertical lines. The center of the coin must be at least 1 cm away from either vertical line and either horizontal line so there is a 1 by 2 rectangular area that it may land in. Thus the probability that there are zero intersections is $\frac{1 \cdot 2}{3 \cdot 4} = \frac{1}{6}$ so the probability there is at least one intersection is $1 - \frac{1}{6} = \frac{5}{6}$.]

[Q2 continued on next page]

(t) Let X be a continuous random variable with probability density function (p.d.f.) $f(x) = 2x$ if $0 \leq x \leq 1$, and $f(x) = 0$ otherwise. Find Var$[X^2]$. *3pts*

$\frac{1}{12}$. [We have Var$[X^2] = \mathbb{E}[X^4] - \mathbb{E}[X^2]^2$. Note that

$$\mathbb{E}[X^4] = \int_0^1 x^4 f(x)dx = \int_0^1 2x^5 dx = \frac{1}{3}$$

and

$$\mathbb{E}[X^2] = \int_0^1 x^2 f(x)dx = \int_0^1 2x^3 dx = \frac{1}{2}$$

so Var$[X^2] = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$.]

(u) Find a and b such that the following function F is a valid c.d.f. for a continuous random variable, and find the corresponding p.d.f. $f(x)$: *3pts*

$$F(x) = \begin{cases} 0, & \text{for } x \leq 0, \\ a(1-x)^2 + b, & \text{for } 0 < x < 1, \\ 1, & \text{for } x \geq 1. \end{cases}$$

$a = -1$; $b = 1$; and $f(x) = 2(1-x)$ for $x \in [0,1]$ (or $(0,1]$, $[0,1)$, $(0,1)$) and $f(x) = 0$ elsewhere. [Since the c.d.f. is continuous, setting $x = 0$ and $x = 1$ in the expression $a(1-x)^2 + b$ should give the values $F(0) = 0$ and $F(1) = 1$, respectively. Thus $a + b = 0$ and $b = 1$. It follows that $a = -1$. Since $f(x) = F'(x)$, then $f(x) = 2(1-x)$ for $x \in [0,1]$ (or $(0,1]$, $[0,1)$, $(0,1)$) and $f(x) = 0$ for other values of x.]

(v) Consider a two-state Markov chain with transition probability matrix $\boldsymbol{P} = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$, where $0 < a, b < 1$. Find the corresponding stationary distribution $\boldsymbol{\pi} = (\pi_1, \pi_2)$. *3pts*

$(\pi_1, \pi_2) = (\frac{b}{a+b}, \frac{a}{a+b})$. [We have $\pi_1 + \pi_2 = 1$ because (π_1, π_2) is a distribution. From the balance equations we have $\pi_1 = (1-a) \cdot \pi_1 + b \cdot \pi_2$ so $a\pi_1 = b\pi_2$. Thus $(1 + \frac{a}{b})\pi_1 = 1$ so $(\pi_1, \pi_2) = (\frac{b}{a+b}, \frac{a}{a+b})$.]

(w) There are two candy jars labeled A and B. Each day, you choose jar A with probability p and jar B with probability $1-p$, and eat one of the candies from the chosen jar. Let $\alpha(i,j)$ denote the probability that jar A becomes empty before jar B, given that jar A currently contains i candies while jar B contains j candies. For $i, j > 0$, write down a recursive formula satisfied by $\alpha(i,j)$: *2pts*

$\alpha(i,j) = p \cdot \alpha(i-1,j) + (1-p) \cdot \alpha(i,j-1)$. [With probability p we pick a candy from jar A and enter a state with $i-1$ jar A candies and j jar B candies. With probability $1-p$ we pick from jar B and have i jar A candies and $j-1$ jar B candies.]

Write down the boundary conditions (base cases) for your recursion: *1pt*

$\alpha(0,j) = 1$ and $\alpha(i,0) = 0$ for $i, j > 0$. [The recursion tells us how to deal with $i, j > 0$ so we need to handle the case when one of them is 0. If there are 0 jar A candies, then jar A has become empty before jar B. If there are 0 jar B candies, then jar A has failed to become empty first.]

3. Trees [*All parts to be justified. Total of 10 pts.*]

(a) Prove by strong induction on the number of vertices that the vertices of any tree can be colored with two colors so that no two adjacent vertices get the same color. [Hint: Remove an arbitrary edge from the tree.] *5pts*

We use strong induction on the number of vertices in tree T. For the base case, T has only a single vertex and can obviously be colored with a single color.

For the induction hypothesis, we assume for all $1 \leq k < n$ that a tree with k vertices can be two-colored.

For the induction step, let T be an arbitrary tree with n vertices. Remove any edge $e = (v_1, v_2)$ from T. Since T is a tree, removing e will disconnect it into two subtrees, T_1 and T_2, each with fewer than n vertices. We can choose the labels of these trees so that $v_1 \in T_1$ and $v_2 \in T_2$.

By the (strong) induction hypothesis, we can two-color T_1 and T_2 separately (using the same set of two colors). Now suppose we add the edge e back in to re-connect the colored trees T_1 and T_2 . If v_1, v_2 were assigned opposite colors, we have a valid two-coloring of T and we are done.

Otherwise, v_1 and v_2 are adjacent vertices with the same color, which is not permitted. To deal with this case, before adding back edge e we can first "flip" the two colors on T_1 so that now v_1, v_2 have opposite colors. Adding edge e again leaves us with a valid two-coloring of T.

Alternative proof: The following induction argument also works (but is not as elegant because it uses the non-trivial fact that any tree has at least one leaf). The base case and induction hypothesis are as above. For the inductive step, let T be an arbitrary tree with n vertices. Using the fact that T has a leaf vertex v, remove the single edge (u, v) connecting v to T, leaving a tree T' with $n - 1$ vertices. Now, by the induction hypothesis, we can color T' with two colors. Finally, color v with the opposite color to that of u and add back the edge (u, v). This gives us a valid two-coloring of the whole of T.

(b) Let T be a tree with n vertices. For $1 \leq i \leq n-1$, let n_i denote the number of vertices in T of degree exactly i. Show that $\sum_{i=1}^{n-1} i n_i = 2(n-1)$. [Reminder: You may use without proof any results from notes or lecture, provided that they are clearly stated.] *2pts*

The left hand side is another way to write $\sum_{v \in V} \deg(v)$, since $\deg(v) = i$ will appear exactly n_i times in $\sum_{v \in V} \deg(v)$. Note that the maximum degree of a vertex in a n-vertex tree is $(n-1)$, while a tree is connected so all vertices must have degree at least 1; this explains why we only need to consider $1 \leq i \leq (n-1)$. The right hand side is $2|E|$, as T is a tree with n vertices, and thus has $(n-1)$ edges. Rewriting the statement, we have $\sum_{v \in V} \deg(v) = 2|E|$, which was proven in class. (It's also easy to see: summing the degrees of all vertices counts each edge exactly twice—once for each of its two endpoints.)

(c) Suppose $n \geq 6$ is even and suppose that T has $\frac{n}{2} + 2$ vertices of degree 1. Prove using part (b) that T must have at least one vertex of degree at least 4. *3pts*

Part (b) tells us $\sum_{i=1}^{n-1} i n_i = 2(n-1)$. Since $\frac{n}{2}+2$ vertices have degree 1, we can apply this information to conclude $\left[\sum_{i=2}^{n-1} i n_i\right] + 1 \cdot \left(\frac{n}{2} + 2\right) = 2(n-1)$. Simplifying further, we have $\sum_{i=2}^{n-1} i n_i = \frac{3n}{2} - 4$.

Now, assume for contradiction that T has no vertices of degree at least 4, so all of its remaining vertices must have degree 2 or 3. The largest possible value for $\sum_{i=2}^{n-1} in_i$ is therefore $3 \cdot \left(\frac{n}{2} - 2\right) = \frac{3n}{2} - 6$, corresponding to all remaining $\left(\frac{n}{2} - 2\right)$ vertices having degree 3. However, by the previous paragraph, the sum is $\frac{3n}{2} - 4$, which is strictly greater. Hence we get a contradiction, so our assumption that T has no vertices of degree at least 4 must be false.

4. Wilson's Theorem [*All parts to be justified. Total of 12 pts.*]

This question leads you through a proof of Wilson's Theorem, which says the following:

Theorem: *A natural number* $n > 1$ *is prime* **if and only if** $(n-1)! \equiv -1 \pmod{n}$.

(a) Assume first that $n > 1$ is not prime, and let $1 < q < n$ be a divisor of n. Show that if a number $a \equiv b \pmod{n}$ then also $a \equiv b \pmod{q}$. *3pts*

If $a \equiv b \pmod{n}$, then $a - b = nm$ for some integer m. Thus $a - b = qkm$ for some integer k since q is a divisor of n. We can conclude that $a - b \equiv 0 \pmod{q}$, i.e., $a \equiv b \pmod{q}$.

(b) Deduce from the previous part that, when $n > 1$ is not prime, $(n-1)! \not\equiv -1 \pmod{n}$. *3pts*

If n is not prime, then n has some divisor q with $1 < q < n$. Assume for the sake of contradiction that $(n-1)! \equiv -1 \pmod{n}$. Then, by part (a), $(n-1)! \equiv -1 \pmod{q}$. But since $1 < q < n$, $(n-1)!$ contains q as a factor, and hence $(n-1)! \equiv 0 \pmod{q}$, a contradiction. So our proof is complete.

(c) Now assume that $n > 1$ is prime. Let x be any number in the range $1 \le x \le n-1$. Show that the only such x which are their own inverse $\pmod{n}$ are $x = 1$ and $x = n - 1 \equiv -1 \pmod{n}$. [Hint: Think about polynomials!] *3pts*

Solution 1: If x is its own inverse modulo n, then $x^2 \equiv 1 \pmod{n}$. To find all such x we want to find the roots of $x^2 - 1 = 0$ over $GF(n)$. This is a degree 2 polynomial over a finite field (because n is prime) so it has at most 2 distinct roots. Since $x = 1$ and $x = -1$ both satisfy the equation, then there can be no other values that are their own inverse modulo n.

Solution 2: If x is its own inverse modulo n, then $x^2 \equiv 1 \pmod{n}$. Thus, there exists some integer k for which $x^2 - 1 = kn$ so n divides $x^2 - 1 = (x-1)(x+1)$. Since n is prime, then n divides one of $x - 1$ or $x + 1$ and the result follows.

(d) Deduce from the previous part that, when $n > 1$ is prime, $(n-1)! \equiv -1 \pmod{n}$. *3pts*

We can write $(n-1)! = (n-1) \cdot (n-2) \cdots (2) \cdot (1)$. By part (c), any integer m with $1 < m < n-1$ is not its own inverse modulo n. Thus, there exists some other integer $m' \neq m$ with $1 < m' < n-1$ such that $m \cdot m' \equiv 1 \pmod{n}$. Since no two distinct elements share the same inverse, we can pair up all the integers strictly between 1 and $n-1$ and deduce that $(n-1)! \equiv (n-1) \cdot (1) \equiv -1 \pmod{n}$ as desired.

5. Partitions via Random Sampling [*No justification necessary. Total of 13 points.*]

There are two urns: Urn 1 has 10 blue and 6 red marbles, while Urn 2 has 7 blue and 9 red marbles. Alice, Bob, and Carol decide to divide up the marbles among them using random sampling. They will first choose one of the urns uniformly at random (u.a.r.), and then use the following scheme to sample from the same chosen urn until it is empty: In each round of sampling, one person is chosen u.a.r., and that person will sample a marble u.a.r. from the urn and keep the marble. Let B_A, B_B, B_C respectively denote the number of blue marbles that Alice, Bob, and Carol have at the end; R_A, R_B, R_C are similarly defined for red marbles.

Whenever possible, express all combinatorial factors in terms of binomial coefficients.

(a) Find $\mathbb{P}[B_A = 3, B_B = 5, B_C = 2 \mid \text{Urn 1 was chosen}]$. *3pts*

$\binom{10}{3}\binom{7}{5}\frac{1}{3^{10}}$. [Each blue marble is equally likely to be drawn by Alice, Bob or Carol. So, given that Urn 1 was chosen, each particular assignment $(i_1, \ldots, i_{10})$, where $i_k \in \{\text{Alice, Bob, Carol}\}$ denotes the person who gets the kth blue marble, has probability $(1/3)^{10}$. There are $\binom{10}{3}\binom{7}{3}$ such assignments that result in 3 blue marbles for Alice, 5 for Bob, and 2 for Carol.]

(b) Find $\mathbb{P}[B_A = 4, R_A = 2 \mid \text{Urn 1 was chosen}]$. *3pts*

$\binom{10}{4}\binom{6}{2}\frac{2^{10}}{3^{16}}$. [Conditioned on Urn 1 being chosen, B_A and R_A are independent, and so

$$\mathbb{P}(B_A = 4, R_A = 2|\text{Urn 1 was chosen}) = \mathbb{P}(B_A = 4|\text{Urn 1 was chosen})\mathbb{P}(R_A = 2|\text{Urn 1 was chosen}).$$

Each factor on the right hand side is a binomial probability: $\mathbb{P}(B_A = 4 \mid \text{Urn 1 was chosen}) = \binom{10}{4}\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)^6$ and $\mathbb{P}(R_A = 2 \mid \text{Urn 1 was chosen}) = \binom{6}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^4$, which follow from the fact that each marble is equally likely to be drawn by Alice, Bob or Carol, so the probability that Alice gets (respectively, does not get) a given marble is $1/3$ (respectively, $2/3$).]

(c) Let E denote the event that 3 blue and 2 red marbles are observed in the first five rounds of sampling. Find $\mathbb{P}[E \mid \text{Urn 1 was chosen}]$ and $\mathbb{P}[\text{Urn 1 was chosen} \mid E]$. *4pts*

$\mathbb{P}(E \mid \text{Urn 1 was chosen}) = \frac{\binom{10}{3}\binom{6}{2}}{\binom{16}{5}} = \frac{\binom{5}{2}\binom{11}{4}}{\binom{16}{6}} = \binom{5}{2}\frac{10\cdot 9\cdot 8\cdot 6\cdot 5}{16\cdot 15\cdot 14\cdot 13\cdot 12}$. [To see the first of these expressions, note that there are $\binom{16}{5}$ ways of choosing a set of five marbles from Urn 1, each of which is equally likely; $\binom{10}{3}\binom{6}{2}$ of those contain exactly 3 blue and 2 red marbles. To see the second expression, note that among all orderings of the 16 marbles in Urn 1, there are $\binom{16}{6}$ ways of picking the positions of the red marbles, all equally likely; of these, $\binom{5}{2}\binom{11}{4}$ have 2 red marbles in the first five positions (and 4 in the remaining 11 positions); to see this last claim, note that once the positions of the red marbles are fixed, those of the blue marbles are also determined. To see the third and final expression, note that there are $\binom{5}{2}$ red-blue sequences consisting of 3 blue and 2 red marbles, each of which has probability given by the stated fraction: to see this, note that the denominator is the total number of ordered sequences of five marbles, while the numerator is the number of such sequences that correspond to a given red-blue sequence.]

$\mathbb{P}(\text{Urn 1 was chosen} \mid E) = \frac{\binom{10}{3}\binom{6}{2}}{\binom{10}{3}\binom{6}{2} + \binom{7}{3}\binom{9}{2}} = \frac{10}{17}$. [Let U_i denote the event that Urn i was chosen. Then, by Bayes' Rule, we have $\mathbb{P}(U_1 \mid E) = \frac{\mathbb{P}(E|U_1)\mathbb{P}(U_1)}{\mathbb{P}(E|U_1)\mathbb{P}(U_1)+\mathbb{P}(E|U_2)\mathbb{P}(U_2)}$. The answer follows from plugging in $\mathbb{P}(U_1) = \mathbb{P}(U_2) = 1/2$, $\mathbb{P}(E \mid U_1)$ from above, and a similar result for $\mathbb{P}(E \mid U_2)$.]

(d) What is the total number of distinct partitions of the marbles among Alice, Bob, and Carol, given that Urn 1 was chosen? [Remark: Marbles of the same color are indistinguishable, but marbles of different colors are distinguishable.] *3pts*

$\binom{12}{2}\binom{8}{2}$. [Partitioning only the blue marbles is a stars and bars problem with $n_B = 10$ stars and $k_B = 2$ bars. Similarly, distributing only the red marbles is a stars and bars problems with $n_R = 6$ stars, $k_R = 2$ bars. Since both choices of partitions are made independently of each other, the total number of partitions is $\binom{n_B+k_B}{k_B} \times \binom{n_R+k_R}{k_R}$.]

6. Probability Bounds and Limits [*No partial credit for (a)-(d). Total of 13 points.*]

Consider i.i.d. random variables $X_1, X_2, \ldots$ with probability distribution $\mathbb{P}[X_i = 2] = \frac{1}{4}$, $\mathbb{P}[X_i = 4] = \frac{1}{2}$, and $\mathbb{P}[X_i = 6] = \frac{1}{4}$ for all $i = 1, 2, \ldots$. Let $S_n = X_1 + X_2 + \cdots + X_n$.

(a) Find $\mathbb{E}[S_n]$ and $\text{Var}[S_n]$. *3pts*

We have $\mathbb{E}[X_i] = 2 \cdot \frac{1}{4} + 4 \cdot \frac{1}{2} + 6 \cdot \frac{1}{4} = 4$, so $\mathbb{E}[S_n] = 4n$ by linearity of expectation. We have $\text{Var}[X_i] = (2-4)^2 \cdot \frac{1}{4} + (6-4)^2 \cdot \frac{1}{4} = 2$, so $\text{Var}[S_n] = 2n$ by independence of $X_1, \ldots, X_n$.

(b) Markov's inequality implies which of the following? Shade only one bubble. *2pts*

- ● $\mathbb{P}[S_n < 5n] \geq \frac{1}{5}$
- ○ $\mathbb{P}[S_n < 5n] \leq \frac{1}{5}$
- ○ $\mathbb{P}[S_n < 5n] \geq \frac{4}{5}$
- ○ $\mathbb{P}[S_n < 5n] \leq \frac{4}{5}$
- ○ None of the above.

Markov's Inequality implies $\mathbb{P}[S_n \geq 5n] \leq \frac{\mathbb{E}[S_n]}{5n} = \frac{4}{5}$, so $\mathbb{P}[S_n < 5n] = 1 - \mathbb{P}[S_n \geq 5n] \geq 1 - \frac{4}{5} = \frac{1}{5}$.

(c) Chebyshev's inequality implies which of the following? Shade only one bubble. *3pts*

- ○ $\mathbb{P}[S_n < 5n] \geq \frac{1}{n}$
- ○ $\mathbb{P}[S_n < 5n] \leq \frac{1}{n}$
- ● $\mathbb{P}[S_n < 5n] \geq 1 - \frac{1}{n}$
- ○ $\mathbb{P}[S_n < 5n] \leq 1 - \frac{1}{n}$
- ○ None of the above.

Chebyshev's inequality tells us that $\mathbb{P}[|S_n - 4n| \geq n] \leq \frac{2n}{n^2} = \frac{2}{n}$. Observe that X_i is symmetrically distributed around 4, which means S_n is symmetrically distributed around $4n$. Thus $\mathbb{P}[S_n \geq 5n] = \mathbb{P}[S_n \leq 3n]$. Since $\mathbb{P}[|S_n - 4n| \geq n] = \mathbb{P}[S_n \geq 5n] + \mathbb{P}[S_n \leq 3n] = 2\mathbb{P}[S_n \geq 5n]$, we have $\mathbb{P}[S_n \geq 5n] \leq \frac{1}{n}$. It then follows that $\mathbb{P}[S_n < 5n] \geq 1 - \frac{1}{n}$.

(d) Let Φ denote the c.d.f. of the standard normal distribution. For large n, which of the following is true? Shade only one bubble. *2pts*

- ○ $\mathbb{P}[S_n \geq 4n + \epsilon n] \approx \Phi(\sqrt{\frac{n}{2}}\,\epsilon)$
- ● $\mathbb{P}[S_n \geq 4n + \epsilon n] \approx 1 - \Phi(\sqrt{\frac{n}{2}}\,\epsilon)$
- ○ $\mathbb{P}[S_n \geq 4n + \epsilon n] \approx \Phi(\frac{\epsilon}{2})$
- ○ $\mathbb{P}[S_n \geq 4n + \epsilon n] \approx 1 - \Phi(\frac{\epsilon}{2})$
- ○ None of the above.

By the Central Limit Theorem, for large n we have $\mathbb{P}\left[S_n \geq 4n + \epsilon n\right] = \mathbb{P}\left[\frac{S_n - 4n}{\sqrt{2n}} \geq \epsilon\sqrt{\frac{n}{2}}\right] \approx 1 - \Phi(\sqrt{\frac{n}{2}}\epsilon)$.

(e) For $\delta > 0$, $\lim_{n\to\infty} \mathbb{P}[S_n \leq (4-\delta)n] = \boxed{0}$. Justify your answer below. *3pts*

This result can be shown in several ways:

1. By the Law of Large Numbers, $\lim_{n\to\infty} \mathbb{P}[S_n \leq (4-\delta)n] = \lim_{n\to\infty} \mathbb{P}\left[\frac{S_n}{n} - 4 \leq -\delta\right] = 0.$
2. A related approach is to use Chebyshev's inequality, which implies
$$\mathbb{P}(S_n - 4n \leq -\delta n) \leq \mathbb{P}(|S_n - 4n| \geq \delta n) \leq \frac{\text{Var}[S_n]}{\delta^2 n^2} = \frac{2n}{\delta^2 n^2} = \frac{2}{\delta^2 n}.$$
The upper bound on the right hand side approaches 0 as $n \to \infty$, so $\lim_{n\to\infty} \mathbb{P}[S_n \leq (4-\delta)n] = 0$.

3. Alternatively, by the Central Limit Theorem, we have $\mathbb{P}\left[S_n \leq 4n - \delta n\right] = \mathbb{P}\left[\frac{S_n - 4n}{\sqrt{2n}} \leq -\sqrt{\frac{n}{2}}\delta\right] \approx \Phi(-\sqrt{\frac{n}{2}}\delta)$ for large n, so $\lim_{n\to\infty} \mathbb{P}[S_n \leq (4-\delta)n] = \lim_{n\to\infty} \Phi\left(-\sqrt{\frac{n}{2}}\delta\right) = 0$.

7. Poisson Distribution [*Justification required where stated. Total of 12 points.*]

Assume that the number of data blocks received at a data storage center per month follows a Poisson distribution with rate $\lambda > 0$, and assume that these numbers over different months are mutually independent. After each month of storage, each data block has probability $p > 0$ of getting corrupted, independently of all other data blocks. Let X_0 denote the number of new data blocks received this month, and, for $n \in \mathbb{Z}^+$, let X_n denote the number of data blocks received n months ago that have so far not been corrupted.

(a) Prove that $X_1 \sim \text{Poisson}[(1-p)\lambda]$. *5pts*

Note that this problem is essentially the same as the Poisson Question discussed in HKN Review Session (see slides 31-32) and further detailed in followup discussions on Piazza.

We need to show that $\mathbb{P}(X_1 = k) = e^{-\lambda_1}\frac{\lambda_1^k}{k!}$ for every $k \in \mathbb{N}$, where $\lambda_1 = (1-p)\lambda$. Let R_1 denote the number of data blocks received one month ago. Then $R_1 \sim \text{Poisson}(\lambda)$ and $X_1 \mid R_1 = r \sim \text{Binomial}(r, (1-p))$. Hence, using the law of total probability, we obtain

$$\begin{aligned}\mathbb{P}(X_1 = k) &= \sum_{r=k}^{\infty}\mathbb{P}(X_1 = k \mid R_1 = r)\mathbb{P}(R_1 = r) \\ &= \sum_{r=k}^{\infty}\binom{r}{k}(1-p)^k p^{r-k} e^{-\lambda}\frac{\lambda^r}{r!} = \frac{(1-p)^k\lambda^k}{k!}e^{-\lambda}\sum_{r=k}^{\infty}\frac{\lambda^{r-k}}{(r-k)!}p^{r-k} \\ &= \frac{[(1-p)\lambda]^k}{k!}e^{-\lambda}\sum_{j=0}^{\infty}\frac{(\lambda p)^j}{j!} = \frac{\lambda_1^k}{k!}e^{-\lambda}e^{\lambda p} = \frac{\lambda_1^k}{k!}e^{-\lambda_1},\end{aligned}$$

where we recognized $\sum_{j=0}^{\infty}\frac{(\lambda p)^j}{j!}$ as the Taylor Series of $e^{\lambda p}$. $\square$

(b) For $n \in \mathbb{Z}^+$, what is the distribution of X_n? State its name and specify its parameter(s). No justification necessary. *2pts*

$\text{Poisson}[(1-p)^n\lambda]$. [If we denote by $R_n \sim \text{Poisson}(\lambda)$ the number of data blocks received n months ago, then by the same reasoning as in part (a), a $\text{Poisson}[(1-p)\lambda]$ number of blocks will remain uncorrupted after the first month. Then, each of this Poisson number of blocks gets corrupted with probability p in the subsequent one month, and so we are left with a $\text{Poisson}[(1-p)(1-p)\lambda] = \text{Poisson}[(1-p)^2\lambda]$ number of uncorrupted blocks after two months. Repeating this reasoning n times, we find $X_n \sim \text{Poisson}[(1-p)^n\lambda]$.

(c) What is the distribution of $X_0 + X_1 + \cdots + X_n$? State its name and specify its parameter(s). No justification necessary. *2pts*

$\text{Poisson}\left[\sum_{k=0}^{n}(1-p)^k\lambda\right]$ or $\text{Poisson}\left[\frac{\lambda}{p}\left(1-(1-p)^{n+1}\right)\right]$. [Since $X_0, \ldots, X_n$ are mutually independent Poisson random variables, part (b) tells us that $X_0 + \cdots + X_n \sim \text{Poisson}(\mu)$, where $\mu = \sum_{k=0}^{n}(1-p)^k\lambda = \frac{\lambda}{p}\left[1-(1-p)^{n+1}\right]$.]

[Q7 continued on next page]

(d) $\lim_{\lambda\to\infty} \mathbb{P}\big[X_0 - \lambda < \sqrt{\lambda}\big] = \Phi(1)$.

[Your answer may be left as an unevaluated sum or integral.] Justify your answer below. *3pts*

For $\lambda \in \mathbb{Z}^+$, we can write $X_0 = Y_1 + \cdots + Y_\lambda$ for $Y_k \overset{\text{i.i.d.}}{\sim} \text{Poisson}(1)$, and since $\mathbb{E}(Y_k) = \text{Var}(Y_k) = 1$, the Central Limit Theorem informs us that

$$\mathbb{P}[X_0 - \lambda < \sqrt{\lambda}] = \mathbb{P}\left[\frac{\sum_{k=1}^{\lambda} Y_k - \lambda}{\sqrt{\lambda}} < 1\right] \to \Phi(1) = \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, \mathrm{d}x$$

as $\lambda \to \infty$.

8. I.I.D. Continuous Uniform Random Variables [*All parts to be justified. Total of 13 points.*]

For $n \geq 2$, let $X_1, \ldots, X_n$ be *independent* Uniform$[0,1]$ random variables, and, for $i \in \{1, \ldots, n\}$, let Y_i denote the ith smallest value of $\{X_1, \ldots, X_n\}$. For example, $Y_1 = \min\{X_1, \ldots, X_n\}$, while $Y_n = \max\{X_1, \ldots, X_n\}$. In HW 12, you found the distributions of Y_1 and Y_n.

(a) Prove that the probability density function (p.d.f.) of Y_2 is given by $f_{Y_2}(y) = n(n-1)y(1-y)^{n-2}$, where $0 \leq y \leq 1$. [Hint: First derive the cumulative distribution function of Y_2.] *4pts*

To compute the p.d.f. of Y_2, we can first compute its c.d.f. and differentiate:

$$\mathbb{P}[Y_2 \leq y] = 1 - \mathbb{P}[Y_2 > y],$$

$$\mathbb{P}[Y_2 > y] = \left(\sum_{i=1}^{n} \mathbb{P}[X_i \leq y \text{ and } X_j > y \text{ for all } j \neq i]\right) + \mathbb{P}[X_i > y \text{ for all } i]$$

$$= \left(\sum_{i=1}^{n} y(1-y)^{n-1}\right) + (1-y)^n = ny(1-y)^{n-1} + (1-y)^n.$$

We get the second line by accounting for two disjoint cases that get us $Y_2 > y$. Either (1) the smallest X_i is less than or equal to y and the rest of the X_i are greater than y, or (2) all X_i are all greater than y. Case (1) has n subcases, as there are n ways to choose which X_i that is less than or equal y.

Now that we have the c.d.f. $F_{Y_2}(y) = 1 - ny(1-y)^{n-1} - (1-y)^n$, we can differentiate it with respect to y to get the p.d.f.:

$$f_{Y_2}(y) = \frac{d}{dy}\left[1 - ny(1-y)^{n-1} - (1-y)^n\right]$$

$$= ny \cdot (n-1)(1-y)^{n-2} - n(1-y)^{n-1} + n(1-y)^{n-1}$$

$$= n(n-1)y(1-y)^{n-2}, \text{ as desired.}$$

(b) For the case of $n = 2$, find the joint p.d.f. $f(y_1, y_2)$ of Y_1 and Y_2. Justify your answer. *3pts*

First, because X_1 and X_2 must be in $[0,1]$, we only care about the domain $0 \leq y_1, y_2 \leq 1$. We know that the joint density $f(y_1, y_2)$ is given as the limit of probabilities of (Y_1, Y_2) being in small rectangles $[y_1, y_1 + dy_1] \times [y_2, y_2 + dy_2]$:

$$f(y_1, y_2) = \lim_{dy_1, dy_2 \to 0} \mathbb{P}(y_1 \leq Y_1 \leq y_1 + dy_1, y_2 \leq Y_2 \leq y_2 + dy_2)/(dy_1 dy_2).$$

This limit is 0 if $y_1 > y_2$ (as Y_1 is necessarily smaller than Y_2). If $y_1 \leq y_2$, since $(Y_1, Y_2) \in [y_1, y_1 + dy_1] \times [y_2, y_2 + dy_2]$ if and only if either $(X_1, X_2) \in [y_1, y_1 + dy_1] \times [y_2, y_2 + dy_2]$ or $(X_2, X_1) \in [y_1, y_1 + dy_1] \times [y_2, y_2 + dy_2]$, we have

$$\mathbb{P}(y_1 \leq Y_1 \leq y_1 + dy_1, y_2 \leq Y_2 \leq y_2 + dy_2) = \mathbb{P}(y_1 \leq X_1 \leq y_1 + dy_1, y_2 \leq X_2 \leq y_2 + dy_2) + \mathbb{P}(y_1 \leq X_2 \leq y_1 + dy_1, y_2 \leq X_1 \leq y_2 + dy_2).$$

Furthermore, since X_1 and X_2 are independent Uniform$[0,1]$ random variables, we have $\mathbb{P}(y_1 \leq X_1 \leq y_1 + dy_1, y_2 \leq X_2 \leq y_2 + dy_2) = \mathbb{P}(y_1 \leq X_2 \leq y_1 + dy_1, y_2 \leq X_1 \leq y_2 + dy_2) \approx dy_1 dy_2$. Hence,

$$f(y_1, y_2) = \begin{cases} 2, & \text{if } 0 \leq y_1 \leq y_2 \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

[Q8 continued on next page]

Alternative Solution: To compute the joint p.d.f. for $0 \le y_1 \le y_2 \le 1$, we can compute the c.d.f. and differentiate it with respect to y_1 and y_2:

$$\begin{aligned}\mathbb{P}[Y_1 \le y_1 \text{ and } Y_2 \le y_2] &= \mathbb{P}[X_1 \le y_1 \text{ and } X_2 \le y_1] \\ &\quad + \mathbb{P}[X_1 \le y_1 \text{ and } y_1 < X_2 \le y_2] + \mathbb{P}[X_2 \le y_1 \text{ and } y_1 < X_1 \le y_2] \\ &= y_1^2 + 2y_1(y_2 - y_1). \\ f(y_1, y_2) &= \frac{d^2}{dy_1 dy_2}[y_1^2 + 2y_1(y_2 - y_1)] = 2.\end{aligned}$$

(c) Assume again $n = 2$ and let $G = Y_2 - Y_1$, the gap size between Y_1 and Y_2. Find the p.d.f. of G. Justify your answer. *3pts*

We can rewrite $G = Y_2 - Y_1$ as $G = |X_2 - X_1|$. Then, the c.d.f. F_G of G is

$$F_G(g) = \mathbb{P}[G \le g] = \mathbb{P}[|X_2 - X_1| \le g] = 1 - \mathbb{P}[|X_2 - X_1| > g],$$

where $\mathbb{P}[|X_2 - X_1| > g]$ can be found as follows. Visualize (X_1, X_2) on the coordinate plane, where the x-axis corresponds to X_1 and the y-axis corresponds to X_2, as shown in Figure 1. Then, $\mathbb{P}[|X_2 - X_1| > g]$ can be found by summing the areas of the two red triangles, which is $(1 - g)^2$.

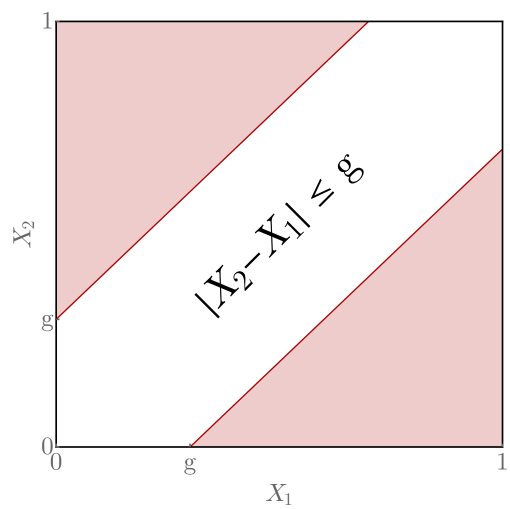


Figure 1: Each red triangle has height and base $(1 - g)$, so the total area of red regions is $(1 - g)^2$.

Now, the p.d.f. of G can be obtained by differentiating the c.d.f. F_G with respect to g:

$$f_G(g) = \frac{d}{dg}\left[1 - (1 - g)^2\right] = 2(1 - g)$$

for $g \in [0, 1]$, and $f_G(g) = 0$ for $g \notin [0, 1]$.

[Q8 continued on next page]

Alternative Solution 1: If you prefer to use calculus, $\mathbb{P}[|X_2 - X_1| > g]$ can also be found as

$$\begin{aligned}
\mathbb{P}[|X_2 - X_1| > g] &= \mathbb{P}[X_2 - X_1 > g] + \mathbb{P}[X_1 - X_2 > g] \\
&= 2 \cdot \mathbb{P}[X_2 - X_1 > g] \quad \text{(by symmetry)} \\
&= 2 \cdot \int_0^{(1-g)} \mathbb{P}[X_2 - X_1 > g | X_1 = x_1] \cdot f_{X_1}(x_1) dx_1 \\
&= 2 \cdot \int_0^{(1-g)} \mathbb{P}[X_2 > g + x_1] \cdot 1 \cdot dx_1 \\
&= 2 \cdot \int_0^{(1-g)} (1 - g - x_1) dx_1 \\
&= 2 \cdot \left[x_1(1-g) - \frac{x_1^2}{2} \right]\Bigg|_0^{(1-g)} = 2 \cdot \left[(1-g)^2 - \frac{(1-g)^2}{2} \right] = (1-g)^2.
\end{aligned}$$

Alternative Solution 2: We can compute the p.d.f. of G using the joint p.d.f. of (Y_1, Y_2) as

$$f_G(g) = \int_0^{1-g} f(y, y+g) dy = 2(1-g),$$

for $g \in [0, 1]$, and $f_G(g) = 0$ for $g \notin [0, 1]$.

(d) What is $\mathbb{P}[G > \frac{1}{2}]$? Justify your answer. [Hint: You should be able to solve this problem without using the p.d.f. of G.] *3pts*

Using $\mathbb{P}[G > g] = \mathbb{P}[|X_2 - X_1| > g] = (1-g)^2$ found above, we get:

$$\mathbb{P}\left[G > \frac{1}{2}\right] = \left(1 - \frac{1}{2}\right)^2 = \frac{1}{4}.$$

[End of Exam]