Chris Piech Fall 2023
CS109

CS109 Final Exam

This is a closed calculator/computer exam. You are, however, allowed to use notes in the exam.

In the event of an incorrect answer, any explanation you provide of how you obtained your answer can potentially allow us to give you partial credit for a problem. For example, describe the distributions and parameter values you used, where appropriate. It is fine for your answers to include summations, products, factorials, exponentials, and combinations.

You can leave your answer in terms of Φ (the CDF of the standard normal) or Φ^{-1} (the inverse CDF). For example $\Phi\left(\frac{3}{4}\right)$ is an acceptable final answer. Recall that the exam is going to be "curved" according to the difficulty of the questions and as such hard questions will not translate to lower grades.

I acknowledge and accept the letter and spirit of the honor code. I pledge to write more neatly than I have in my entire life:

Signature: _____

Family Name (print): _____

Given Name (print): _____

Email (preferably your gradescope email): _____

1 Short Answer (40 points)

Answer each of the following questions. You must give a **brief** justification for your answer.

a. (8 points) Let $Z \sim \text{Exp}(\lambda = 1)$ be the minutes until someone coughs during the exam. What is the probability that someone coughs within the next 2 mins?

$$\begin{aligned} P(Z < 2) &= F(2) \\ &= 1 - e^{-\lambda \cdot 2} \\ &= 1 - e^{-2} \approx 0.86 \end{aligned}$$

b. (7 points) What is the probability that the `random_sum` function returns a value greater than 501?

```
def random_sum():
 total = 0
 for i in range(1000):
   total += random() # sample from a uniform(0, 1)
 return total
```

Let X be the sum of 1000 uniform. By the CLT, this is well approximated by a normal distribution $Y \sim N(\mu = 1000 \cdot 0.5, \sigma^2 = 1000 \cdot \frac{1}{12})$. You do not need a continuity correction because X and Y are both continuous.

$$\begin{aligned} P(X > 501) &\approx P(Y > 501) \\ &= 1 - P(Y < 501) \\ &= 1 - \phi\left(\frac{501 - 500}{\sqrt{1000/12}}\right) \end{aligned}$$

c. (8 points) Google weather says there is an overall 10% chance of the event, R, that it will rain at some point today. They also divide the day into 8 time intervals and provide the probability for the event that it rains during each of those intervals: $W_1, W_2, \ldots W_8$. Based on only these probabilities how could you test if the eight events are (a) mutually exclusive (b) independent. Leave your answers to both (a) and (b) as expressions in terms of $P(R)$ and $P(W_1) \ldots P(W_8)$.

If the events are mutually exclusive, this expression must hold:

$$P(R) = \sum_i P(W_i)$$

If the events are independent, this expression must hold:

$$\begin{aligned} P(R) &= 1 - P(R^C) \\ &= 1 - \prod_{i=1}^{8} P(W_i^C) \\ &= 1 - \prod_{i=1}^{8} 1 - P(W_i) \end{aligned}$$

d. (5 points) Let $X = A + B + C + D + E + F$ where A through F are all independent random variables. Does the Central Limit Theorem apply to X? Explain as if teaching in 1 or 2 sentences.

$$\begin{aligned} A &\sim \text{Bernoulli}(p = 0.5) \\ B &\sim \text{Binomial}(n = 4, p = 0.8) \\ C &\sim \text{Geometric}(p = 0.3) \\ D &\sim \text{Uni}(\alpha = 0, \beta = 1) \\ E &\sim \text{Beta}(a = 2, b = 3) \\ F &\sim \text{Exp}(\lambda = 3) \end{aligned}$$

No! They are independent, but not identically distributed.

e. (5 points) In order to make optimal parking decisions you need to represent your belief in X, the probability that parking spaces on a particular road will be open (parking spaces are either open or not). Before observations you believe the probability that a space is open is Uniform($\alpha = 0, \beta = 1$). You pass 10 parking spots on the road, 9 of which are full, one of which was open. What is your posterior belief in X if you consider the 10 spots to be IID samples?

$$X \sim \text{Beta}(a = 2, b = 10)$$

A uniform prior between 0 and 1 is also a Beta($a = 1$, $b = 1$). We observe 1 success (open parking spot) and 9 failures (full parking spots), so we update our Beta parameters by adding 1 to a and 9 to b.

f. (7 points) In Google's recent release of Gemini (their ChatGPT competitor) they reported that during the training process, "silent data corruption" affected the integrity of their training data. Silent data corruption refers to random bit flips in a computer's data, potentially leading to flawed outcomes or system malfunctions. They have estimated that they experience an average of 0.1 instances of silent data corruption per terabyte of data processed. If the training process for their next model requires processing 50 terabytes of data, calculate the probability that there will be less than two instances of silent data corruption during training.

Let X be the number of SDCs. On average there are $50 \cdot 0.1 = 5$ instances of SDC per 50tb.

$$X \sim \text{Poi}(\lambda = 5)$$

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) \\ &= \frac{5^0 e^{-5}}{0!} + \frac{5^1 e^{-5}}{1!} \\ &= e^{-5} + 5e^{-5} \\ &= 6e^{-5} \end{aligned}$$

2 Don't Quit Until You Are Ahead (22 points)

A player is playing uno (a card game with no ties). Each game the player has a 0.55 probability of winning. Each game is independent.

a. (7 points) The player plays 10 games. What is the exact probability they have more wins than losses?

Let X be the number of games they win. $X \sim \text{Bin}(n = 10, p = 0.55)$.

$$\begin{aligned} P(X > 5) &= \sum_{i=6}^{10} P(X = i) \\ &= \sum_{i=6}^{10} \binom{10}{i} \cdot 0.55^i \cdot 0.45^{10-i} \end{aligned}$$

b. (7 points) The player plays i games (where $i > 20$). What is an approximation for the probability that they have more wins than losses? Assume that i is even.

Let Y be an approximation for X. By the CLT, $Y \sim N(\mu = i \cdot 0.55, \sigma^2 = i \cdot 0.55 \cdot 0.45)$. If i is even, having more wins than losses means having $i/2 + 1$ or more wins. The $(i + 1)/2$ below includes continuity correction (adding $1/2$).

$$\begin{aligned} P(X > i/2) &\approx P(Y > (i + 1)/2) \\ &= 1 - \phi\left(\frac{(i + 1)/2 - \mu}{\sigma}\right) \end{aligned}$$

c. (8 points) Let's try a different angle! Instead of playing a fixed number of games, the player stops **as soon as** they have more wins than losses. What is the probability they stop after exactly 5 games?

There are only two cases where this is true: [LWLWW] and [LLWWW]. Let E be the event that you stop after exactly 5 games.

$$\begin{aligned} P(E) &= P(LWLWW) + P(LLWWW) \\ &= 2 \cdot 0.45^2 \cdot 0.55^3 \end{aligned}$$

3 Popcorn (23 points)

The instructions on a popcorn bag say to keep cooking until there is 1 second between pops. A "pop" is the sound made when a popcorn kernel turns into a fluffy piece of popcorn.

The time in seconds since you started cooking for a popcorn kernel to pop is well approximated as $T_i \sim N(\mu = 100, \sigma^2 = 20^2)$. There are 100 kernels in a microwavable bag. Aside: even though T_i is "time until", it is not Exponential as it doesn't follow the Poisson process. Each kernel's popping time is IID.

a. (7 points) 110 seconds have passed and 99 kernels have popped. What is the probability that the final unpopped kernel will pop in the next 1 second?

$$\begin{aligned} P(T_i < 111 | T_i > 110) &= \frac{P(T_i < 111 \text{ and } T_i > 110))}{P(T_i > 110)} \\ &= \frac{P(T_i < 111) - P(T_i < 110)}{1 - P(T_i < 110)} \\ &= \frac{\phi(\frac{111-100}{20}) - \phi(\frac{110-100}{20})}{1 - \phi(\frac{110-100}{20})} \\ &= \frac{\phi(\frac{11}{20}) - \phi(\frac{1}{2})}{1 - \phi(\frac{1}{2})} = \frac{0.017378}{0.308538} \approx 0.056 \end{aligned}$$

I expect many people will solve this to be $P(110 < T_i < 111)$ which is reasonable, but isn't correct, because we know that it didn't pop in the first 110 seconds.

b. (5 points) 110 seconds have passed and 70 kernels have popped. What is the probability that at least one of the 30 remaining kernels will pop in the next 1 second? Let p_a be your answer to part (a).

Every event in this answer is conditioned on 110 seconds passing. Let E be the event that at least one kernel pops in the next 1 second. Let K_i be the event that kernel i pops in the next 1 second.

$$\begin{aligned} P(E) &= 1 - \prod_{i=1}^{30} P(K_i^C) \\ &= 1 - \prod_{i=1}^{30} (1 - p_a) \\ &= 1 - (1 - p_a)^{30} \end{aligned}$$

I expect many people will solve this to be $P(110 < T_i < 111)$ which is reasonable, but isn't correct, because we know that it didn't pop in the first 110 seconds.

c. (10 points) 110 seconds have passed and an unknown number of kernels have popped. What is the probability that at least one of the remaining kernels will pop in the next 1 second?

The probability that a single kernel pops in the first 110s is

$$P(T_i < 110) = \phi(\frac{110 - 100}{20}) = \phi(\frac{1}{2})$$

Let N be the number of popped kernels after 110s. $N \sim \text{Bin}(100, p = \phi(\frac{1}{2}))$.

If there are $100-n$ kernels left, the probability of at least one pop in the next 1 second is $1-(1-p_a)^{100-n}$

Let E be the event of a pop in the next one second. We can solve for $P(E)$ using the law of total probability:

$$\begin{aligned} P(E) &= \sum_{n=0}^{100} P(E|N=n)P(N=n) \\ &= \sum_{n=0}^{100} \left(1-(1-p_a)^{100-n}\right) \cdot \binom{100}{n} \cdot \phi(\frac{1}{2})^n \cdot (1-\phi(\frac{1}{2}))^{100-n} \end{aligned}$$

The result is the same if the sum is from 0 to 99, but it must start at 0, not 1.

4 Delicate Polling (25 points)

You are trying to estimate the probability p that a randomly selected person in a population thinks the answer to a single delicate true or false question is "true". However the topic is sensitive, so if you ask them directly, they will not provide an honest answer.

Instead of directly asking the question, we are going to give each participant 7 true or false questions and ask them to report **how many** they think are true. Question 1 is the "delicate" question we care about. Each of the other 6 questions are unimportant, nonsensitive questions where we know that each person is equally likely to answer "true" of "false", independent of their response to other questions.

a. (8 points) Let X_i be the number of "true" answers for person i. What is the probability that $X_i = 4$? Leave your answer in terms of p, the probability that a person thinks the answer to question 1 is true.

Let Z_i be an indicator variable for whether the person thinks question 1 is true. Let $Y_i \sim \text{Bin}(n = 6, p = 0.5)$ be the count of the other six the agree with.

$$\begin{aligned} P(X_i = 4) &= P(Y_i = 3, Z_i = 1) + P(Y_i = 4, Z_i = 0) \\ &= P(Y_i = 3)P(Z_i = 1) + P(Y_i = 4)P(Z_i) = 0 \\ &= p\binom{6}{3}0.5^6 + (1-p)\binom{6}{4}0.5^6 \end{aligned}$$

b. (8 points) Write an expression for $E[X_i]$ in terms of p.

Note that $X_i = Y_i + Z_i$. Which makes it easy to compute the expectation.

$$\begin{aligned} E[X_i] &= E[Y_i] + E[Z_i] \\ &= 6 \cdot 0.5 + p \\ &= 3 + p \end{aligned}$$

c. (9 points) Your solution to part (b) can be expressed as $E[X_i] = \alpha p + \beta$ where α and β are computed constants. Now we want to estimate p from a sample of 100 people, where we observe X_i for each person. To do this, we use the function `estimate_p`, which estimates $E[X_i]$ to be the sample mean and then chooses a value for p such that $E[X_i] = \alpha p + \beta$. Write bootstrap pseudocode to approximate and `print` the probability that `estimate_p(samples)` is within 0.1 of the actual fraction of the population that believes the answer to question 1 is 'true'.

```
def estimate_p(samples):
    # np.mean calculates the average of a list
    sample_mean = np.mean(samples)
    return (sample_mean - beta)/alpha

def bootstrap_p_distribution(samples):
    estimated_p = estimate_p(samples)
    # your code here:
```

```
n_tests = 10000
n_within = 0
for i in range n_tests:
→ resamples = np.random.choice(samples, len(samples), replace=True)
→ reestimate = estimate_p(resamples)
→ diff = abs(reestimate - estimated_p)
→ if diff < 0.1:
→→ n_within += 1
print(n_within / n_tests)
```

5 Water Levels (20 points)

The fraction of water in a reservoir at any point in time is modelled by a ReservoirDistribution with a single parameter, a. The ReservoirDistribution has the following PDF and CDF:

$$\text{PDF: } f(X = x) = 2ax^{a-1}(1 - x^a)$$

$$\text{CDF: } F(x) = 1 - (1 - x^a)^2$$

a. (6 points) For a particular reservoir, at a particular time of the year, we estimate $a = 1.5$. What is the probability that the reservoir is more than 0.25 full at that time of the year?

$$\begin{aligned} P(X > 0.25) &= 1 - F(0.25) \\ &= 1 - (1 - (1 - 0.25^{1.5})^2) \\ &= (1 - 0.25^{1.5})^2 \end{aligned}$$

b. (14 points) For a new reservoir, you observe n measurements of fullness: $x_1, x_2, \ldots, x_n$. Explain, in words, how you would choose parameter a using the maximum likelihood estimation framework, and provide any necessary derivatives. Note: $\frac{d}{dx}k^x = k^x \log(k)$

$$\begin{aligned} L(a) &= \prod_i f(x_i) \\ &= \prod_i 2ax_i^{a-1}(1 - x_i^a) \\ LL(a) &= \sum_i \log 2 + \log a + (a-1)\log x_i + \log(1 - x_i^a) \\ \frac{\partial LL(a)}{\partial a} &= \sum_i \frac{1}{a} + \log x_i + \frac{1}{1 - x_i^a}\frac{\partial}{\partial a}(1 - x_i^a) \\ &= \sum_i \frac{1}{a} + \log x_i - \frac{x_i^a \log x_i}{1 - x_i^a} \end{aligned}$$

6 Logistic Regression Priors (28 points)

We are going to train a logistic regression model – but we don't have very much training data. As such, we would like to use MAP, instead of MLE (the parameter estimation method we used on pset6). We are going to use a simplified logistic regression model that has a single parameter θ, for a single feature x. Specifically we assume: $P(Y = 1|X = x) = \sigma(\theta \cdot x)$ where x and θ are each numbers, not lists, and there is no bias term. Our prior belief is that θ is standard normal, $N(\mu = 0, \sigma = 1)$. Our training data is a list of n tuples: $[(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)]$

a. (8 points) Based **only** on your prior belief of θ, how many times larger is the probability density that $\theta = 0$, compared to the probability density that $\theta = 2$?

$$\frac{f(\theta=0)}{f(\theta=2)} = \frac{\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{0-\mu}{\sigma}\right)^2}}{\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{2-\mu}{\sigma}\right)^2}} = \frac{e^{-\frac{1}{2}\left(\frac{0-0}{1}\right)^2}}{e^{-\frac{1}{2}\left(\frac{2-0}{1}\right)^2}} = \frac{1}{e^{-2}} = e^2$$

b. (10 points) The MAP objective for this simplified logistic regression is to choose the θ value that is most likely given the values of y in our training data: $\theta_{\text{MAP}} = \underset{\theta}{\text{argmax}} f(\theta|y_1, \ldots, y_n)$. Write out and explain, step by step, how to derive:

$$\theta_{\text{MAP}} = \underset{\theta}{\text{argmax}} \left[\log(f(\theta)) + \sum_{i=1}^{n} \log(f(y_i|\theta))\right]$$

$$\begin{aligned}
\theta_{\text{MAP}} &= \underset{\theta}{\text{argmax}} f(\theta|y_1, \ldots, y_n) \\
&= \underset{\theta}{\text{argmax}} f(y_1, \ldots, y_n|\theta) f(\theta) && \text{Bayes Theorem} \\
&= \underset{\theta}{\text{argmax}} \prod_i f(y_i|\theta) f(\theta) && \text{Since data are IID} \\
&= \underset{\theta}{\text{argmax}} \log f(\theta) + \sum_i \log f(y_i|\theta) && \text{argmax of log}
\end{aligned}$$

c. (10 points) We want to apply gradient ascent to find θ. Write an expression that calculates the gradient we should use. Recall:

$$\theta_{\text{MAP}} = \underset{\theta}{\text{argmax}} \left[\log(f(\theta)) + \sum_{i=1}^{n} \log(f(y_i|\theta))\right]$$

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\Big[\log(f(\theta)) + \sum_{i=1}^{n}\log(f(y_i|\theta))\Big] \\
&\quad= \frac{\partial}{\partial\theta}\Big[\log\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\theta^2} + \sum_i \log\left(\sigma(\theta\cdot x_i)^{y_i}(1-\sigma(\theta\cdot x_i))^{(1-y_i)}\right)\Big] \\
&\quad= \frac{\partial}{\partial\theta}\Big[-\frac{1}{2}\theta^2 + \sum_i y_i\log\sigma(\theta\cdot x_i) + (1-y_i)\log(1-\sigma(\theta\cdot x_i))\Big] \\
&\quad= -\theta + \sum_i y_i\frac{1}{\sigma(\theta\cdot x_i)}\sigma(\theta\cdot x_i)(1-\sigma(\theta\cdot x_i))x_i + (1-y_i)\frac{-1}{1-\sigma(\theta\cdot x_i)}\sigma(\theta\cdot x_i)(1-\sigma(\theta\cdot x_i))x_i \\
&\quad= -\theta + \sum_i y_i(1-\sigma(\theta\cdot x_i))x_i - (1-y_i)\sigma(\theta\cdot x_i)x_i \\
&\quad= -\theta + \sum_i y_i x_i - y_i\sigma(\theta\cdot x_i))x_i - \sigma(\theta\cdot x_i)x_i + y_i\sigma(\theta\cdot x_i)x_i \\
&\quad= -\theta + \sum_i (y_i - \sigma(\theta\cdot x_i))x_i
\end{aligned}$$

7 Estimating Child Vocab Size (22 points)

Aside: while this question is secretly about estimating vocab size in children, we are going to use dice, as it is easier to think about during an exam.

A child is rolling an n-sided fair dice (the sides have the integers 1 through n, and each side is equally likely). The child rolls the dice three times and gets the value 12 twice and 48 once. Clearly the dice has at least 48 sides, but it could have more! Your prior belief is that each value of n between 1 and 100 (including 1 and 100) is equally likely.

a. (8 points) If the dice has 50 sides, what is the probability of rolling 12 twice and 48 once, in any order?

We can either use the multinomial, or, derive it from first principles. Let X_i be the count of outcome i:

$$P(X_{12} = 2, X_{48} = 1) = \binom{3}{2, 1}\left(\frac{1}{50}\right)^3$$
$$= \frac{3}{50^3}$$

b. (14 points) What is a mathematical expression for the probability that the dice has n sides, given that the child has rolled 12 twice and 48 once?

Note the probability is 0 if $n < 48$. Assuming $48 \leq n \leq 100$:
Let N be the number of sides on the dice. Let D be the event where we observe $X_{12} = 2, X_{48} = 1$.

$$P(D|N = n) = \frac{3}{n^3}$$

$$P(N = n) = \frac{1}{100}$$

$$P(N = n|D) = \frac{P(D|N = n)P(N = n)}{\sum_{i=1}^{100} P(D|N = i)P(N = i)} \quad \text{Bayes}$$
$$= \frac{\frac{3}{n^3}\frac{1}{100}}{\sum_{i=48}^{100} \frac{3}{i^3}\frac{1}{100}} \quad \text{substitute}$$
$$= \frac{1}{n^3 \cdot \sum_{i=48}^{100} \frac{1}{i^3}} \quad \text{simplify}$$

That's all folks. Thank you all for the wonderful quarter and we hope you have a fantastic winter break!