Chris Piech
CS109

Fall 2023

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?

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
```

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 $P(W_1) \ldots P(W_8)$.

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.

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

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?

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.

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?

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.

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?

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?

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).

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?

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.

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

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:
```

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?

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: $\frac{d}{dx} k^x = k^x \log(k)$

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 $\theta = 2$?

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:

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

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

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

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?

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?

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