Groups: This pset must be done individually. You are encouraged to discuss problem-solving strategies with other classmates as well as the course staff, but you must write up your own solutions.

Instructions: For each problem, remember you must briefly explain/justify how you obtained your answer, as correct answers without an explanation will receive no credit. Moreover, in the event of an incorrect answer, we can still try to give you partial credit based on the explanation you provide. It is fine for your answers to include summations, products, factorials, exponentials, or combinations; you don’t need to calculate those all out to get a single numeric answer.

Submission: You must upload your written compiled LaTeX PDF to Gradescope under “PSet 5 [Written]” and your code files bootstrap.py and mab.py to “PSet 5 [Coding]”. You must tag your written problems on Gradescope, or you will receive no credit as mentioned in the syllabus. Please cite any collaboration at the top of your submission.

Note that no late days are allowed on this problem set, unlike prior problem sets. There will also not be an “on-time bonus” since everyone has to submit on-time.

1. Let \( x = (x_1, \ldots, x_n) \) be iid samples from the density function

\[
f_X(x; \theta) = \begin{cases} 
\theta x^{\theta - 1} & 0 \leq x \leq 1 \\
0 & \text{otherwise}
\end{cases}
\]

where \( \theta \) is an unknown, but fixed parameter of interest (not a random variable).

(a) What is the maximum likelihood estimator, \( \hat{\theta}_{MLE} \) of \( \theta \)? Also, verify that your estimator is indeed a maximizer (and not a minimizer).

(b) What is the method of moments estimator, \( \hat{\theta}_{MoM} \) of \( \theta \)? You may write your answer in term of the sample mean \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \).

(c) Suppose the observed samples were \( x = (0.43, 0.62, 0.99, 0.98) \). What are your maximum likelihood and method of moments estimators for \( \theta \)? Give your answers to four decimal places.

2. Let \( x = (x_1, \ldots, x_n) \) be iid samples from \( Pois(\Theta) \) where \( \Theta \) is a random variable (not fixed).

(a) Using the prior \( \Theta \sim Gamma(r, \lambda) \) (for some arbitrary but known parameters \( r, \lambda > 0 \)), show that the posterior distribution \( \Theta|x \) also follows a Gamma distribution and identify its parameters (by computing \( \pi_\Theta(\theta|x) \)). Then, explain this sentence: “The Gamma distribution is the conjugate prior for the rate parameter of the Poisson distribution”. Hint: This can be done in just a few lines!
(b) Now derive the MAP estimate for $\Theta$. The mode of a $\text{Gamma}(s, \nu)$ distribution is $\frac{s - 1}{\nu}$. Hint: This should be just one line using your answer to part (a).

(c) Explain how this MAP estimate differs from the MLE estimate (recall for the Poisson distribution it was just the sample mean $\frac{\sum_{i=1}^{n} x_i}{n}$, see 7.2 notes), and provide an interpretation of $r$ and $\lambda$ as to how they affect the estimate.

(d) (Extra Credit): Suppose $n = 254, \bar{x} = 2.1, r = 7, \lambda = 12$. Construct a 96% credible interval (watch/read 8.2) for $\Theta$, and give your answer to six decimal places.

3. Let $x = (x_1, \ldots, x_n)$ be iid samples from $\text{Exp}(\theta)$ where $\theta$ is unknown but fixed (not a RV). Recall both the MLE/MoM estimates for $\theta$ were $\hat{\theta} = \frac{1}{\bar{x}}$ where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ is the sample mean. We will construct a 98% confidence interval for $\theta$ using the following procedure. (You need not refer to the confidence interval notes for any formulae - we provide each step below.)

(a) Recall from PSet4 that if $V \sim \text{Gamma}(r, \lambda)$, then $M_V(t) = (\frac{\lambda}{\lambda-t})^r$. Use properties of the MGF to show that $aV \sim \text{Gamma}(r, \lambda/\alpha)$ for any scalar $\alpha > 0$.

(b) Using your answer to part (a), what is the (exact) distribution of the quantity $n\bar{x}\theta$ (now treating $\bar{x}$ as a random variable)? Hint: $n\bar{x} = \sum_{i=1}^{n} x_i$ is just the sum of the $n$ iid $\text{Exp}(\theta)$ random variables. Your parameter(s) should NOT depend on $\theta$.

(c) Let $n = 263$. Find the values of $a, b$ such that $P(a \leq n\bar{x}\theta \leq b) = 0.98$ - there are infinitely many that work, but choose the ones so that $P(n\bar{x}\theta \leq a) = 0.01$ and $P(n\bar{x}\theta \leq b) = 0.99$ to get a symmetric one. To find the point $t$ such that $F_T(t) = y$ for $T \sim \text{Gamma}(u, v)$, call the following function which gets the inverse CDF:

$$\text{scipy.stats.gamma.ppf}(y, u, \theta, 1/v)$$

(The third parameter of 0 is weird, but must be set. You can ignore it). You can run your code in this special slide on Edstem, but no need to turn this in. Give your answers for $a, b$ rounded to six decimal places.

(d) Finally, rearrange the equation above, plugging in your answers from (c), to get a 98% confidence interval for $\theta$, assuming $\bar{x} = 0.134$. Give your answer in the form $[c, d]$ where $c, d$ are rounded to six decimal places.

(e) Instead of finding this exact distribution of $n\bar{x}\theta$ (and hence finding an exact confidence interval), we could have used the CLT instead to approximate its distribution! Repeat the process above (starting with part (b)) using the CLT to find the approximate distribution of $n\bar{x}\theta$, reporting your approximate confidence interval to six decimal places. Justify in one-two sentences why we may use the CLT. (Hint: Look up the mean and variance of the RV from part (b), and use a Normal approximation.)

(f) (Extra Credit): Remember we just chose the two endpoints in part (c) for symmetry even though there were infinitely many intervals which contain 98% probability (e.g., $P(n\bar{x}\theta \leq a') = 0.015$ and $P(n\bar{x}\theta \leq b') = 0.995$), and hence may not have minimum length. Write code to find the (exact) 98% confidence interval $[a', b']$ that is as narrow as possible (minimizing $b' - a'$). Attach your code here using the verbatim environment, and report your confidence interval rounded to three decimal places.
4. You have recently taken over the management of admissions to The GaussHouse, a mansion for content creators. The house specializes in producing educational short videos about CS109 content. Prospective residents have to apply and be selected to live in GaussHouse. Admission should be based on ability to make CS109 educational content, but is it?

The previous manager left behind data from thousands of applications from last year. You are curious about whether the selection process was biased towards Generation Z (people under 24). In addition to the age category, you also have access to whether or not an applicant had previously produced and uploaded a probability related short video, which we will call “Content” for short.

<table>
<thead>
<tr>
<th>Gen-Z</th>
<th>Supp. Video</th>
<th>Accepted</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>.003</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>.0001</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>.0007</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>.0112</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.026</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>.2705</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>.9994</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>.5891</td>
</tr>
</tbody>
</table>

(a) What is the probability that an applicant is Gen-Z? **Give your answer to four decimal places.**

(b) Given that an applicant was accepted, what is the probability that the applicant was Gen-Z? Given that an applicant was accepted, what is the probability that the applicant was not Gen-Z? **Give your answers to four decimal places.**

(c) Next let us calculate a few fairness tests. Recall that in our class on fairness we defined random variables $A$ for a demographic variable, and $R$ for result. $R = 1$ is a positive result, in this case acceptance.

i. **Independence Fairness** means that acceptance rates should be the same for all groups. It is defined as:

$$ P(R = 1|A = a) = P(R = 1|A = b) \quad (1) $$

Did the previous admissions to GaussHouse satisfy independence with respect to age?

ii. **Relaxed Independence Condition** is defined as:

$$ \frac{P(R = 1|A = a)}{P(R = 1|A = b)} \geq 1 - \varepsilon \quad (2) $$

Where $A = a$ is the event that an applicant is from the group which is less represented in the applicant pool and $A = b$ is the event that an applicant is from the more represented group. Did previous admissions satisfy relaxed independence with respect to age for $\varepsilon = .2$?
(d) Your house wants to update their admission system, and one of their goals is to satisfy “fairness through unawareness”. One option is to hide an applicant’s age from people who review applications. You notice that applications still have people’s names. US census data makes it very clear that names change substantially in popularity over time. How could the inclusion of an applicant’s name impact fairness through unawareness even if age is removed?

(e) You want to ensure that future admissions to GaussHouse are fair. Which of the types of fairness that we talked about in class (fairness through unawareness; independence; relaxed independence; separation) would you choose and why? In your answer you should explain in words what each of the four types of fairness means in this context. Base your answer on what you think would be the best fairness standard for this context, not on what current US discrimination law requires. A precise, well justified opinion will receive full credit.
The following two questions are cumulative (may draw from any lecture).

5. You are on your way to buy tickets to see the new hit probability movie, “μ Girls” (get it?) with your friends! At the movie theater, you have to make the usual decision - which line should I wait in? The lines are long so you and your friends briefly observe the rate of people served per minute in each line. How should we balance the rate at which people are served and the length of a line? The movie is starting soon. You will make it on time if you get your tickets within the next 15 minutes.

<table>
<thead>
<tr>
<th>Line</th>
<th>Rate (people / min)</th>
<th>Num People in line</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line A</td>
<td>1/2</td>
<td>4</td>
</tr>
<tr>
<td>Line B</td>
<td>1/3</td>
<td>3</td>
</tr>
<tr>
<td>Line C</td>
<td>1/4</td>
<td>2</td>
</tr>
</tbody>
</table>

(a) If you choose a line uniformly at random to wait in together, what is the probability that you and your friends are on time for the movie (treat your group as one person who will buy tickets for everyone)? Be careful with the units of time. Give your answer to four decimal places.

(b) Your genius friend says, “why don’t we just each take a line and if anyone is able to buy tickets in time, we can make the movie.” Under her plan, what is the probability that you and your friends are on time for the movie? Assume that each line acts independently. Give your answer to four decimal places.

(c) You and your friends decide to make it interesting - whoever gets to the front will actually not get reimbursed by the other two (i.e., pay for all three tickets). Hence, you do not want to be the one to reach the front of your line first. Let \(X, Y, Z\) be the time until the person in line A, B, and C reaches the front of their line, respectively (at which point, the other two leave their lines happily). Identify from our zoo what the distribution of \(X, Y\) and \(Z\) are (with parameter(s)), and write an expression for the probability that the person in line A pays (for example, you may write something like: \(P(X < Y + Z) - P(\max\{X, Y, Z\} > 5)\)). Ignore the 15-minute constraint for this problem - you will wait indefinitely until someone gets to the front of the line.

(d) Following the same scheme from part (c), suppose the three of you each choose a random line (with each of the 3! assignments equally likely). What is the probability that you pay? Hint: You don’t actually need to do any computation! You may argue your answer to this question in one-two sentences. Give your answer to four decimal places.
6. The final exam in CS 109 consists of 10 pages, and each student must upload exactly 10 images: one for each of the 10 pages in the exam. Careless students

- Shuffle the 10 images randomly, with each possible ordering equally likely.
- Rotate each image randomly (and independently of other images), with each of the 4 rotations equally likely. Only one of these rotations is upright.

We say a single image is **perfect** if it has perfect position (it is in the correct position in the exam), AND has perfect orientation (is upright).

On the other hand, careful students make sure all ten of their images are perfect. (It is possible though unlikely that a careless student has all ten perfect pages as well.)

(a) What is the expected number of perfect images for a careless student? Hint: The range of this RV is \{0, 1, 2, \ldots, 9, 10\}. **Give your answer to four decimal places.**

(b) What is the variance of the number of perfect images for a careless student? **Give your answer to four decimal places.**

(c) Alex thinks that the number of careless students (out of 70) is equally likely to be any integer in \{0, 1, 2, \ldots, 70\} (the remaining students are careful). What is the expected number of perfect images (out of the 700 total images uploaded)? Hint: Be very careful of your expression for this quantity, and remember that a careful student submits 10 perfect images. **Give your answer to four decimal places.**

(d) The deadline has passed, and Alex knows now that exactly 50 students were careless (the remaining 20 students were careful). Compute the probability that at least 220 images are perfect, using the CLT. **Give your answer to four decimal places.**
7. **[Coding+Written]** Suppose you are working at Coursera on new ways of teaching a concept in probability. You have two different learning activities activity1 and activity2 and you want to figure out which activity leads to better learning outcomes. Over a two-week period, you randomly assign each student to be given either activity1 or activity2. You then evaluate each student’s learning outcomes by asking them to solve a set of problems.

You are given iid samples $x_1, \ldots, x_n$ which measure the performance of $n$ students who were given activity1, and iid samples $y_1, \ldots, y_m$ which measure the performance of $m$ students who were given activity2.

The data you are given has the following statistics:

<table>
<thead>
<tr>
<th>Activity</th>
<th>Number of Samples</th>
<th>Sample Mean</th>
<th>Sample Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>activity1</td>
<td>$n = 542$</td>
<td>$\bar{x} = 144.928044$</td>
<td>$s_x^2 = 3496.339840$</td>
</tr>
<tr>
<td>activity2</td>
<td>$m = 510$</td>
<td>$\bar{y} = 153.129412$</td>
<td>$s_y^2 = 5271.763645$</td>
</tr>
</tbody>
</table>

(a) Perform a single hypothesis test using the procedure in 8.4 at the $\alpha = 0.05$ significance level, and report the exact p-value (to four decimal places) for the observed difference in means. In other words: assuming that the learning outcomes for students who had been given activity1 and activity2 had the same mean $\mu_x = \mu_y$, what is the probability that you could have sampled two groups of students such that you could have observed a difference of means as extreme, or more extreme, than the one observed? (Hint: Use the CLT and closure properties of the Normal distribution to compute the distribution of $\bar{X} - \bar{Y}$. What is $\mu_x - \mu_y$ (under the null hypothesis) and what are the variances $\sigma_x^2$, $\sigma_y^2$ (estimates of these are given) of some sample $x_i$ and $y_j$ respectively?)

(b) Now, write code to estimate the p-value using the bootstrap method, instead of computing it exactly. Implement the function `bootstrap_pval` in `bootstrap.py`. Your answer to this part and the previous should be very close! What is your computed p-value (to four decimal places)?
8. [Coding+Written] Suppose you are a data scientist at Facebook and are trying to recommend to your boss Mark Zuckerberg whether or not to release the new PYMK (“People You May Know”) recommender system. They need to determine whether or not making this change will have a positive and **statistically significant** (commonly abbreviated “stat-sig”) impact on a core metric, such as time spent or number of posts viewed.

Facebook could do a standard hypothesis test (called an “A/B Test” in industry), where we compare the same metric across the “A” group (“current system”, the “control group”) vs the “B” group (“new system”, the “experimental group”). If the “B” group has a stat-sig improvement in this metric over the “A” group, we should replace the current system with the new one!

This typically involves putting 99% of the population (Facebook users) in the “A” group, and 1% of the population (1% of 2 billion users is still 20 million users) in the “B” group. This heavily imbalanced distribution has the following consequences:

- If there is an unforeseen negative impact, it doesn’t affect too many people.
- If there is an unforeseen positive impact, it won’t be released as early (loss of tons of possible revenue).

Facebook decides to ditch A/B Testing and try the Multi-Armed (Bernoulli) Bandit approach! There are $K = 2$ arms (whether to use the current system or the new system), and the rewards are Bernoulli: 1 if a user sends (at least) one friend request to someone in PYMK (within a small timeframe of seeing the recommendations), and 0 otherwise. This may not seem like it has impact on revenue, but: more friends $\rightarrow$ more engagement/time spent on FB $\rightarrow$ more ads being shown $\rightarrow$ more revenue.

You will first implement the Upper Confidence Bound and Thompson Sampling algorithms generically before applying it to this Facebook example in the last two parts.

(a) Implement the function `upper_conf_bound` in `mab.py`, following the pseudocode for the UCB1 algorithm. Include here in the writeup the two plots that were generated automatically.

(b) Implement the function `thompson_sampling` in `mab.py`, following the pseudocode for the Thompson Sampling algorithm. Include here in the writeup the two plots that were generated automatically.

(c) Explain in your own words, for each of these algorithms, how both exploration and exploitation were incorporated. Then, analyze the plots - which algorithm do you think did “better” and why?

(d) Suppose Facebook has 500,000 users (so that you can actually run your code in finite time, but they actually have a lot more), and the current recommender system has a true rate of $p_1 = 0.47$ (proportion of users who send (at least) one request), and the new one has a true rate of $p_2 = 0.55$. That is, the new system is actually better than the old one.

- If we performed an A/B Test with 99% of the population in group A (the current system), and only 1% of the population in group B (the new system), what is the expected number of people (out of 500,000) that will send (at least) one friend request?
• If we used the Thompson Sampling algorithm to decide between the two arms (group A and group B), what is the experimental number of people (out of 500,000) that will send (at least) one friend request? (Modify the main function of your code after submitting it to Gradescope. You may also want/need to comment out the call to UCB).

(e) Repeat the previous part but now assume $p_1 = 0.47$ and $p_2 = 0.21$. That is, the new system is actually much worse than the old one. Then, explain in a few sentences the relationships between the 4 numbers produced (2 from this part and 2 from the previous part).

9. (Required) [Survey]: By Monday of the last week, the course evaluations for CS109 will be available on Axess. Please fill it out completely, and attach a screenshot below showing you’ve completed the evaluation for both Tim and Alex. It will be anonymous, so please be honest (constructive criticism is appreciated!). Your feedback may help shape future quarters of CS 109. You all did an amazing job getting through an extremely difficult class during an extremely difficult quarter. You should all be proud of yourselves for making it through - congratulations from the course staff!

10. (Extra Credit) [Coding] Implement Logistic Regression for binary input/output data. Specifically, you should implement the gradient ascent algorithm.

(a) Implement the function fit in logistic_regression.py
(b) Implement the function predict in logistic_regression.py