Data 102, Spring 2023 Midterm 1

- You have 110 minutes to complete this exam. There are 5 questions, totaling 40 points.
- You may use one 8.5×11 sheet of handwritten notes (front and back), and the provided reference sheet. No other notes or resources are allowed.
- You should write your solutions inside this exam sheet.
- You should write your name and Student ID on every sheet (in the provided blanks).
- Make sure to write clearly. We can't give you credit if we can't read your solutions.
- Even if you are unsure about your answer, it is better to write down partial solutions so we can give you partial credit.
- We have provided two blank pages of scratch paper, one at the beginning and one at the end of the exam. No work on these pages will be graded.
- You may, without proof, use theorems and facts that were given in the discussions or lectures, **but please cite them**.
- We don't answer questions individually. If you believe something is unclear, bring your question to us and if we find your question valid we will make a note to the whole class.
- Unless otherwise stated, no work or explanations will be graded for multiple-choice questions.
- Unless otherwise stated, you must show your work for free-response questions in order to receive credit.

Last name	
First name	
Student ID (SID) number	
Berkeley email	
Name of person to your left	
Name of person to your right	

Honor Code

I will respect my classmates and the integrity of this exam by following this honor code.

I affirm:

- All of the work submitted here is my original work.
- I did not collaborate with anyone else on this exam.

Signature: _____

1. (5 points) For each of the following, determine whether the statement is true or false. For this question, no work will be graded and no partial credit will be assigned.

   (a) (1 point) The choice of the constant M or proposal distribution $f(x)$ in rejection sampling has no effect on sampling efficiency, as long as $Mq(x) \leq f(x)$, where $q(x)$ is the unnormalized target density.

   ◯ A. TRUE ◯ B. FALSE

   (b) (1 point) Consider two medical tests for a disease. The first test has TPR=0.9 and FPR=0.05, while the second test has TPR=0.54 and FPR=0.03. Then the first test will always have a higher FDR than the second test.

   ◯ A. TRUE ◯ B. FALSE

   (c) (1 point) Given specific sample data, the Benjamini-Hochberg procedure guarantees that the FDP will be lower than the requested level α.

   ◯ A. TRUE ◯ B. FALSE

   (d) (1 point) When using a GLM to fit continuous y with the Bernoulli likelihood, the Identity and Sigmoid are valid choices for the inverse link function.

   ◯ A. TRUE ◯ B. FALSE

   (e) (1 point) When conducting linear regression in Bayesian perspective, the choice of prior will determine the form of regularization applied to the model.

   ◯ A. TRUE ◯ B. FALSE

Name: _____ SID: _____

2. (8 points) **A different approach to FWER control.** Consider the following algorithm, known as the Holm-Bonferroni procedure:

   1. Given a significance level $\alpha \in [0, 1]$ and a set of n p-values, $p_1, \ldots, p_n$. Sort the p-values in non-decreasing order: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$
   2. For $k \in \{1, 2, \ldots, n\}$, if $p_{(k)} \leq \frac{\alpha}{n-k+1}$, reject the corresponding null hypothesis and continue. Otherwise, fail to reject all remaining hypotheses.

   (a) (2 points) **For this part only**, we consider the followintg 5 p-values for multiple hypothesis testing:

p-value	threshold	decision	reality
0.001			1
0.007			1
0.01			0
0.1			0
0.16			0

   Fill in the threshold and decision columns of the above table for the Holm-Bonferroni procedure with level $\alpha = 0.05$. How many tests does the procedure reject?

Solution: We reject 3 tests.

p-value	threshold	decision	reality
0.001	0.01	1	1
0.007	0.0125	1	1
0.01	0.0167	1	0
0.1	0.025	0	0
0.16	0.05	0	0

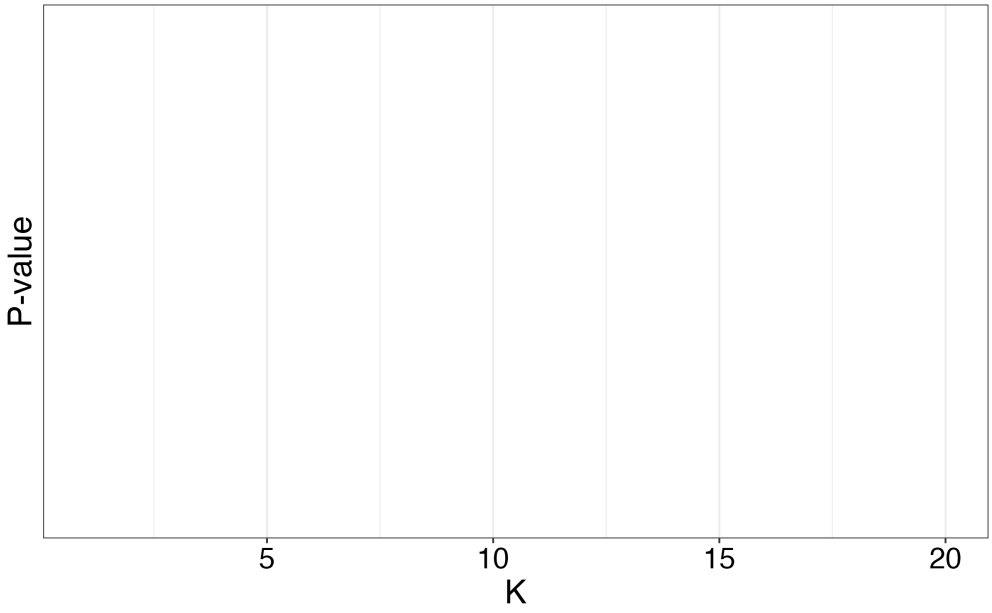
   (b) (1 point) Like the Bonferroni correction, the Holm-Bonferroni procedure controls the family-wise error rate at level α. Does the Holm-Bonferroni method make more or less discoveries than the Bonferroni correction? Justify your answer.

Solution: It is less conservative, since it does not conduct all tests at level $\frac{\alpha}{n}$. Only the first p-value is tested at that level and other tests are made at levels greater than $\frac{\alpha}{n}$.

(c) (1 point) Comparing the Benjamini-Hochberg procedure with Bonferroni, which one makes fewer discoveries? In other words, does Bonferroni at rate α also controls FDR at rate α or Benjamini-Hochberg at rate α also controls FWER at rate α? Justify your answer in words.

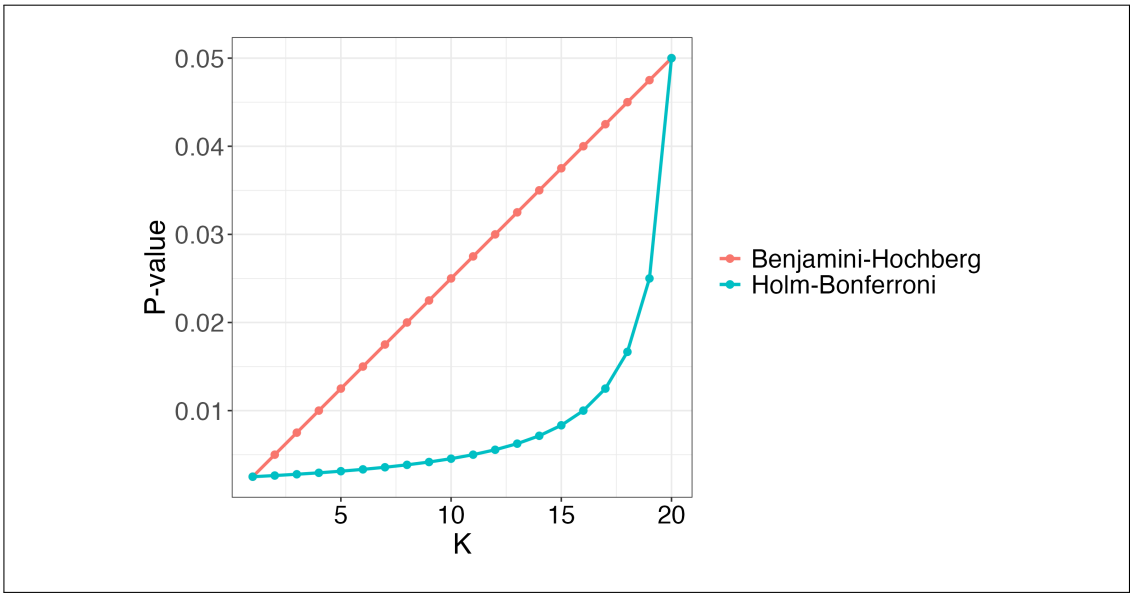
Solution: Bonferroni is more conservative because it always uses a smaller p-value threshold: $\frac{\alpha}{n} \leq \frac{k\alpha}{n}$ for all values $k \in \{1, \ldots, n\}$

(d) (2 points) Assuming $n = 20$, draw the Benjamini-Hochberg guide line and Holm-Bonferroni guide line ($\frac{\alpha}{n-k+1}$) on the same plot. X-axis should be k and the y-axis should be the p-value threshold. The Holm-Bonferroni guideline does not need to be exact, but its shape and position relative to the BH line should be accurate. Make sure you specify the equation for each line.



Solution: BH guideline is $p = \frac{\alpha}{20}k$ whereas HB guideline is $p = \frac{\alpha}{21-k}$
Figure below shows the two guidelines for when $\alpha = 0.05$

Name: _____ SID: _____



(e) (2 points) Comparing the Benjamini-Hochberg procedure with Holm-Bonferroni, which one makes more discoveries for the same significance level α? You must show your work: show *mathematically* that either all discoveries made by Benjamini-Hochberg will also be made by Holm-Bonferroni, or the opposite.

Solution: HB is more conservative. Looking at the plot from previous part, we can see that threshold for Benjamini-Hochberg is always greater than or equal to the threshold for Holm-Bonferroni for $k \in \{1, \ldots, 20\}$. So if HB rejects a tests, BH will definitely reject it as well since it will use a higher p-value threshold on the same p-value.

We can show this mathematically. Let $p_{BH,k}$ and $p_{HB,k}$ be the threshold for the BH and HB procedures on the k^{th} p-value respectively. First we show that $\forall k \in \{1, \ldots, 20\} : p_{HB,k} \leq p_{BH,k}$:

$$\begin{aligned}
&\frac{\alpha}{21-k} \leq \frac{\alpha}{20}k \\
\iff &\frac{1}{21-k} \leq \frac{1}{20}k \\
\iff &21-k \geq \frac{20}{k} \\
\iff &21k - k^2 \geq 20 \\
\iff &k^2 - 21k + 20 \leq 0 \\
\iff &(k-1)(k-20) \leq 0
\end{aligned}$$

The last statement is true because $k \in \{1, \ldots, 20\}$. Now let's suppose $p_{(k^*)}$ is

the last p-value we reject with the Holm-Bonferroni procedure. We have that:

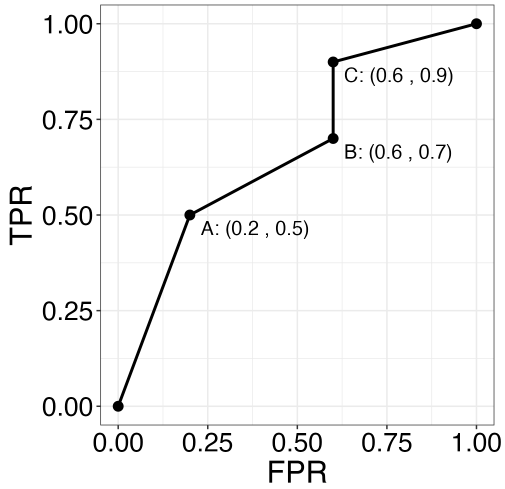
$$p_{(k)} \leq p_{HB,k} \quad \forall k \in \{1, \ldots, k^*\}$$

because we reject all the first k^* test with Holm-Bonferroni. Using the result above we have:

$$p_{(k)} \leq p_{BH,k} \quad \forall k \in \{1, \ldots, k^*\}$$

since $p_{HB,k} \leq p_{BH,k}$. This means if we reject the first k^* test with Holm-Bonferroni, we also reject them, and possibly more, with the Benjamini-Hochberg procedure. So BH is less conservative and makes more discoveries.

3. (9 points) Your friend has developed a new cancer detection algorithm based on imaging and plans to evaluate its performance using an ROC curve. After testing the algorithm on samples from many patients, your friend generates the following ROC curve with the important points labeled with their corresponding (FPR, TPR). Throughout, assume that a positive case corresponds to having cancer.



(a) (1 point) What are the FNR and TNR associated with point B?

Solution: FNR = 0.3 and TNR=0.4

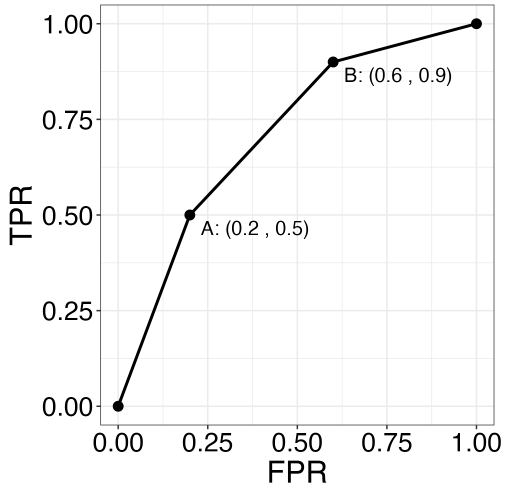
(b) (2 points) Fill in the blanks below and explain your answer.

Among the points A, B, and C, _____ is strictly better than _____.

Solution: C is strictly better than B. Because for the same FPR value, it has a strictly higher TPR value.

(c) (3 points) It is possible to modify the algorithm and obtain the following ROC curve instead. First, explain why the modified algorithm is better: in other words, explain what measure we can use to make such a comparison. Second, describe how we can obtain the improved ROC curve using the algorithm which generated the ROC curve in parts (a), (b).

Solution: The modified algorithm has a higher AUC and its ROC is strictly above the original ROC. In other words, $TPR_{modified} \geq TPR_{original}$ at all values of FPR with strictly inequality at some FPR values.
One can obtain the modified ROC by ignoring all thresholds between points A and B and instead using the decision from point A with probability p or point C with probability $(1-p)$. Changing the value of p will lead to the ROC line



segment between points A and C. The TPR and FPR on that line segment has the following form for $p \in [0, 1]$:

$$\text{TPR} = p * \text{TPR}_A + (1 - p) * \text{TPR}_C$$
$$\text{FPR} = p * \text{FPR}_A + (1 - p) * \text{FPR}_C$$

(d) (3 points) A hospital looks into using this algorithm, and determines the cost of incorrectly classifying a cancer patient as not having cancer is \$1000, whereas the cost of incorrectly classifying a non-cancer patient as having cancer \$100. What should the baseline prevalence of cancer be such that you are indifferent between points A and B in the modified ROC curve from part (c)?

Solution: Let $P(cancer) = \pi$

$$\pi \times 0.5 \times 1000 + (1 - \pi) \times 0.2 \times 100 = \pi \times 0.1 \times 1000 + (1 - \pi) \times 0.6 \times 100$$
$$5\pi + 0.2(1 - \pi) = \pi + 0.6(1 - \pi)$$
$$4\pi = 0.4(1 - \pi)$$
$$\pi = \frac{1}{11} \approx 0.091$$

4. (10 points) Joe has landed a summer internship job in a customer service department. His job is to model the number of complaints received per week. He obtains data corresponding to n weeks $\{x_1, x_2, \ldots, x_n\}$ where x_i is the number of complaints received in week i. Each x_i follows a Poisson distribution with parameter λ:

$$x_i \sim \text{Poisson}(\lambda)$$

The PMF of a Poisson random variable with parameter λ is: $P(x = k) = \frac{\lambda^k e^{-\lambda}}{k!}$. You can assume that the number of complaints received in different weeks are independent of each other: $x_i \perp\!\!\!\perp x_j \; \forall i \neq j$.

(a) (2 points) Suppose he wants to conduct the following hypothesis test:

$$H_0 : \lambda = \lambda_1$$
$$H_1 : \lambda = \lambda_2$$

where $\lambda_2 > \lambda_1$. He needs to fix his significance level at α. Find the decision rule of the most powerful test for his problem. Your decision rule should fill in the blank in the following sentence with a mathematical expression that depends on $x_1, \ldots, x_n$, $\alpha, \lambda_1, \lambda_2$ and other constants: "If _____, then reject the null hypothesis".
Hint: you don't need to simplify your expression or solve for the exact rejection threshold: you can just express it as a function that depends on the constant(s). Make sure you specify which constant(s) affects the threshold.

Solution:

$$\begin{aligned}\Lambda &= \frac{P(x;\lambda_1)}{P(x;\lambda_2)} \\ &= \prod_{i=1}^{n} \frac{\frac{\lambda_1^{x_i} e^{-\lambda_1}}{x_i!}}{\frac{\lambda_2^{x_i} e^{-\lambda_2}}{x_i!}} \\ &= \prod_{i=1}^{n} (\frac{\lambda_1}{\lambda_2})^{x_i} e^{\lambda_2 - \lambda_1}\end{aligned}$$

Reject if $\Lambda > \eta$

where η is the rejection threshold which depends on λ_1 and more importantly significance level α. It is easier to work with log likelihood ratio:

$$\log \Lambda = \sum_{i=1}^{n} x_i \log(\frac{\lambda_1}{\lambda_2}) + (\lambda_2 - \lambda_1)$$

$$\text{Reject if } \sum_{i=1}^{n} x_i > \frac{\log(\eta) + n(\lambda_2 - \lambda_1)}{\log(\lambda_2) - \log(\lambda_1)}$$

where constant η depends on λ_1 and more importantly α

(b) (2 points) Derive the Maximum Likelihood Estimator for λ.

Solution:

$$L(X_1, \ldots, X_N;\ \lambda) = \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} = \frac{\lambda^{\sum_{i=1}^n X_i} e^{-n\lambda}}{\prod_{i=1}^n X_i!}$$

$$\ell(X_1, \ldots, X_N;\ \lambda) = \sum_{i=1}^{n} X_i \log \lambda - n\lambda - \sum_{i=1}^{n} \log X_i!$$

$$\frac{\partial}{\partial \lambda} \ell(X_1, \ldots, X_N;\ \lambda) = \frac{\sum_{i=1}^n X_i}{\lambda} - n$$

$$\frac{\sum_{i=1}^n X_i}{\hat{\lambda}_{MLE}} - n = 0$$

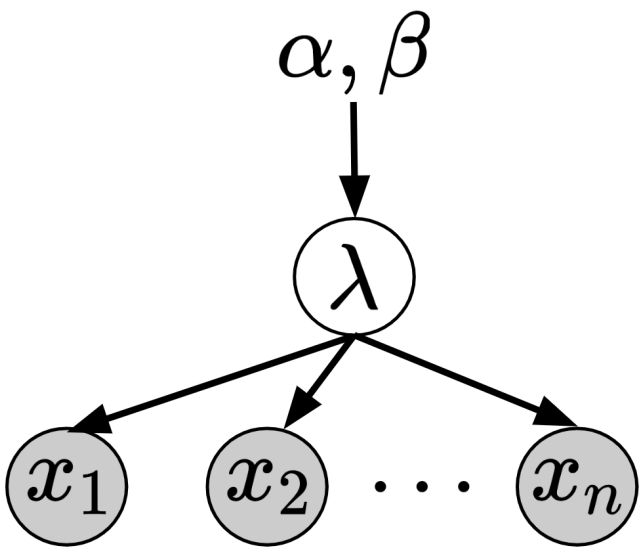
$$\implies \hat{\lambda}_{MLE} = \frac{\sum_{i=1}^n X_i}{n}$$

(c) (1 point) Joe asks for your help to set up the problem from a Bayesian perspective. He makes the following choices:

- λ is distributed according to a Gamma distribution with parameters $\alpha > 0$ and $\beta > 0$: $\lambda \sim \text{Gamma}(\alpha, \beta)$.
- Each x_i is still distributed according to a Poisson with parameter λ and is conditionally independent of other x_j's given λ.

Draw the graphical model for the setup above.

Solution:



(d) (2 points) Using the Bayesian model in the previous part, derive the posterior distribution for λ after observing $\{x_1, \ldots, x_n\}$. The PDF of a Gamma distribution with parameters α and β has the following form:

$$p(y; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}$$

where $\Gamma(\alpha)$ is a function that depends on α which is a constant here. The reference sheet includes more information on the Gamma distribution.
Hint: you should work with the unnormalized posterior which is proportional to the true posterior. You don't need to carry over the constants.

Solution:

$$\lambda | \{x_1, \ldots, x_n\} \sim \text{Gamma}(\alpha + \sum_{i=1}^{n} x_i, \beta + n)$$

In other words, Gamma distribution is the conjugate prior for a Poisson likelihood.

(e) (2 points) Using the posterior distribution from the previous part, find the MAP and MMSE estimate of λ.

Solution: MAP is the mode of the Gamma posterior and MMSE is its expected value.

$$\hat{\lambda}_{MAP} = \frac{\alpha + \sum_{i=1}^{n} x_i - 1}{\beta + n}$$

$$\hat{\lambda}_{MMSE} = \frac{\alpha + \sum_{i=1}^{n} x_i}{\beta + n}$$

(f) (1 point) Compare the MMSE estimate from previous part with the MLE from above. Are they identical? If not, when would they become identical?

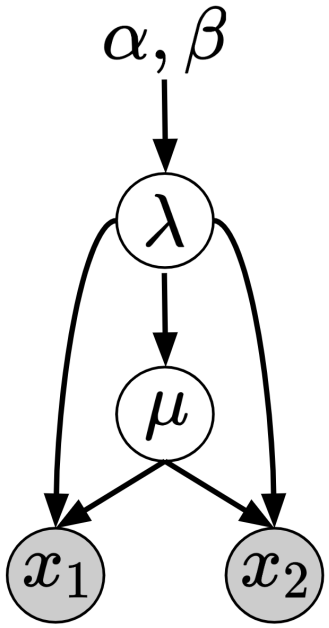
Solution: The prior makes the two estimates different, but as we get more samples ($n \to \infty$) the weight of prior becomes smaller and $\hat{\lambda}_{MMSE} \to \hat{\lambda}_{MLE}$. This makes sense as data size grows, the weight of prior on our posterior becomes less and less. In the limit case where data grows infinitely, prior has no impact and our MAP decision is solely based on Likelihood which depends solely on data. In that case maximizing the posterior (MAP) is the same as maximizing the likelihood (MLE).

5. (8 points) Consider the following model with unknown variables λ and μ, observed variables $x_1, \ldots, x_n$ and known constants α and β

$$\begin{aligned} \lambda &\sim \text{Gamma}(\alpha, \beta) \\ \mu \mid \lambda &\sim \text{Exponential}(\lambda) \\ x_i \mid \mu, \lambda &\sim \mathcal{N}(\mu, \lambda^2) \end{aligned}$$

(a) (1 point) For this part only, suppose $n = 2$. Draw a graphical model for the variables described above.

Solution:



(b) (3 points) The following pseudocode provides a description of Gibbs sampling, but it contains exactly two mistakes. Circle each mistake, and in the space below, write the correct version (if the correction is just to remove the circled part, just write "remove only")

Hint: the fixes involve removing or changing part of the algorithm: no moving around is necessary.

(a) Compute the distributions $p(\lambda|\mu)$ and $p(\mu|\lambda, x_1, \ldots, x_n)$
(b) Initialize the following variables to zero: $\lambda, \mu, x_1, \ldots, x_n$
(c) Repeat the following steps until enough samples have been obtained:
  (i) Using the current values of λ and $x_1, \ldots, x_n$, draw a sample for μ from the conditional distribution in step (a).
  (ii) Using the current values of μ and $x_1, \ldots, x_n$, draw a sample for λ from the other conditional distribution in step (a).
  (iii) Save the current values of λ and μ as samples.

Solution: First mistake: in part (a), conditional distribution of λ should also depend on the data: $p(\lambda|\mu, x_1, \ldots, x_n)$
Second mistake: in part (b), we don't change or initialize the values of data we received. We keep them fixed throughtout the whole process and only initialize the unnknown variables: "Initialize the following variables to zero: λ, μ". Ideally, we could also "initialize λ, μ to random values, positive for λ", not zero necessarily.

(c) (2 points) In step c-(ii) of the Gibbs sampling procedure above, we need to obtain a new sample for λ. Choose the single most efficient sampling algorithm to use to approximate the distribution for λ, or if sampling is not necessary, select option C. Ensure to justify your answer.

◯ A. Rejection sampling
◯ B. Metropolis-Hastings
◯ C. Sampling is not necessary

Solution: The posterior of λ conditioned on μ and x's has no closed form solution mainly because λ is dependent on both variables at the same time. We need to use a sampling method to approximate the posterior. Metropolis-Hasting is usually a better choice since it is more efficient with higher acceptance rate.

(d) (2 points) Suppose we notice we have made a mistake in our model and the variance of each x is actually a known constant σ^2. In other words, our new model is:

$$\begin{aligned} \lambda &\sim \text{Gamma}(\alpha, \beta) \\ \mu \mid \lambda &\sim \text{Exponential}(\lambda) \\ x_i \mid \mu &\sim \mathcal{N}(\mu, \sigma^2) \end{aligned}$$

with unknown variables λ and μ, observed variables $x_1, \ldots, x_n$ and known constants α, β and σ^2.
Now under this new model, what is the most efficient algorithm for sampling λ in step c-(ii) of the Gibbs sampling procedure above? Ensure to justify your answer.

◯ A. Rejection sampling
◯ B. Metropolis-Hastings
◯ C. Sampling is not necessary

Solution: Now the link between X and λ is broken. The posterior of λ only depends on μ. But Gamma distribution is the conjugate prior for Exponential likelihood of μ. So $P(\lambda|\mu)$ is also Gamma distributed with known parameters. Thus, the distribution of λ in step c-(i) is exactly known and there is no need for approximate sampling.

Midterm 1 Reference Sheet

Algorithm 1 The Benjamini-Hochberg Procedure

input: FDR level α, set of n p-values $P_1, \ldots, P_n$

Sort the p-values $P_1, \ldots, P_n$ in non-decreasing order $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(n)}$

Find $K = \max\{i \in \{1, \ldots, n\} : P_{(i)} \leq \frac{\alpha}{n} i\}$

Reject the null hypotheses (declare discoveries) corresponding to $P_{(1)}, \ldots, P_{(K)}$

Useful Distributions:

Distribution	Support	PDF/PMF	Mean	Variance	Mode
$X \sim \text{Poisson}(\lambda)$	$k = 0, 1, 2, \ldots$	$\frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ	$\lfloor \lambda \rfloor$
$X \sim \text{Gamma}(\alpha, \beta)$	$x \geq 0$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\frac{\alpha-1}{\beta}$
$X \sim \mathcal{N}(\mu, \sigma^2)$	$x \in \mathbb{R}$	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$	μ	σ^2	μ
$X \sim \text{Exponential}(\lambda)$	$x \geq 0$	$\lambda \exp(-\lambda x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	0

Conjugate Priors: For observations x_i, $i = 1, \ldots, n$:

Likelihood	Prior	Posterior
$x_i \mid \theta \sim \text{Bernoulli}(\theta)$	$\theta \sim \text{Beta}(\alpha, \beta)$	$\theta \mid x_{1:n} \sim \text{Beta}\left(\alpha + \sum_i x_i, \beta + \sum_i (1 - x_i)\right)$
$x_i \mid \mu \sim \mathcal{N}(\mu, \sigma^2)$	$\mu \sim \mathcal{N}(\mu_0, 1)$	$\mu \mid x_{1:n} \sim \mathcal{N}\left(\frac{\sigma^2}{\sigma^2+n}\left(\mu_0 + \frac{1}{\sigma^2}\sum_i x_i\right), \frac{\sigma^2}{\sigma^2+n}\right)$
$x_i \mid \lambda \sim \text{Exponential}(\lambda)$	$\lambda \sim \text{Gamma}(\alpha, \beta)$	$\lambda \mid x_{1:n} \sim \text{Gamma}\left(\alpha + n, \beta + \sum_i x_i\right)$

Generalized Linear Models

Regression	Inverse link function	Likelihood
Linear	identity	Gaussian
Logistic	sigmoid	Bernoulli
Poisson	exponential	Poisson
Negative binomial	exponential	Negative binomial

Sigmoid function: $\sigma(x) = \dfrac{1}{1 + e^{-x}}$