Name:

Roll No.: **Dept.:**

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: September 19, 2022

Instructions: *Total:* **60 marks**

1. Total duration: **2 hours**. Please write your name, roll number, department on **all pages**.
2. This booklet has 8 pages (6 pages + 2 pages for rough work). No part of your answers should be on pages designated for rough work. Additional rough sheets may be provided if needed.
3. Write/mark your answers clearly in the provided space. Please keep your answers precise and concise.
4. Avoid showing very detailed derivations (you may use the rough sheet for that). In some cases, you may directly use the standard results/expressions provided on page 6 of this booklet.

Section 1 (6 Multiple Choice Questions: 6 x 2 = 12 marks). (Tick/circle all options that you think are true)

1. Which of the following quantities express the model (epistemic) uncertainty? (1) Prior, (2) Posterior, (3) Likelihood, (4) Marginal likelihood
2. Computing which of these quantities, in general, require computing an integral/sum? (1) Expected CLL (2) Marginal likelihood, (3) Log likelihood, (4) Posterior predictive distribution (PPD).
3. For Bayesian linear regression with Gaussian likelihood, (1) The full posterior of the model's unknowns (weight vector and hyperparameters), in general, can be computed exactly, (2) The posterior predictive distribution of this model, in general, can be computed exactly, (3) Assuming Gaussian prior, the conditional posterior (CP) of the weight vector if the hyperparameters are known/fixed can be computed exactly; (4) The mean of the CP of the weight vector is the same as its MLE solution.
4. Assuming a model with Gaussian likelihood $p(X|\theta)$, a general Gaussian prior over a parameter θ, and all other hyperparameters as fixed: (1) The prior promotes θ to take small values, (2) The posterior $p(\theta|X)$ is guaranteed to be Gaussian, (3) The posterior's (whatever it is) variance will keep on increasing as we use more and more training data, (4) The MAP estimation problem for θ will have a unique solution.
5. Which of the following is true about Gaussian Process (GP) regression? (1) The PPD has a closed form expression if the hyperparameters are fixed/known; (2) Cost of computing the PPD scales in the number of training examples; (3) When hyperparameters are fixed, computing the PPD for this model does not require computing the posterior of the GP function; (4) If using the same kernel with same hyperparameters for both GP regression and kernel ridge regression, there is no additional benefit offered by GP regression.
6. Which of the following is true about the Laplace approximation? (1) It is not suitable to approximate distributions that do not have a symmetric shape; (2) It is not suitable to approximate distributions that have multiple modes; (3) Using Laplace approximation for a posterior ensures that the posterior predictive will have a closed form expression; (4) The cost of computing the Laplace approximation scales in the number of parameters.

Section 2 (6 short answer questions: 6 x 3 = 18 marks). .

1. Briefly explain what is the difference between the full posterior and a conditional posterior (CP)? You may use the example of the Bayesian linear regression model (or any other suitable model you wish) to explain the difference.

Name:

Roll No.: **Dept.**:

2. Is the true posterior predictive distribution (PPD) of a Bayesian model equivalent to an ensemble of a finite number of members? Justify your answer briefly.

3. For a Bayesian logistic regression model, suppose the posterior of the weights has been approximated by a Gaussian. Briefly explain how would you approximate the posterior predictive distribution $p(y_*|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y})$, and also write down the expression for $p(y_* = 1|\boldsymbol{x}_*, \mathbf{X}, \boldsymbol{y})$ when using this approximation.

4. In a latent variable model with both global and local variabless, why is it okay/reasonable if we just compute point estimates of the global variables as opposed to their posteriors?

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: September 19, 2022

5. Consider two models $\mathcal{M}_1$ and $\mathcal{M}_2$ for some data $\mathbf{X}$. Suppose you want to decide which of the two is the better model by (1) Comparing their posterior probabilities; and (2) Comparing their marginal likelihoods. Will (1) and (2) give the same result? Briefly justify your answer (maximum 1-2 sentences).

6. Briefly describe what is the intuitive meaning of the hyperparameters of the prior distribution in a probabilistic model. You may use an example to explain.

Section 3 (5 not-so-short answer questions: 5 x 6 = 30 marks). .

1. Consider the Bayesian linear regression model with Gaussian likelihood. Assume locally conjugate priors on the model's unknowns (the weight vector $\boldsymbol{w}$, and the likelihood's and the prior's hyperparameters β and λ). Give a sketch/pseudo-code of the Gibbs sampler to compute the approximate joint posterior $p(\boldsymbol{w}, \beta, \lambda | \mathbf{X}, \boldsymbol{y})$. Clearly mention the conditional posteriors that you will need for this Gibbs sampler. You do not have to derive or write down the final expressions of these CPs; only specify the CPs you will need.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: September 19, 2022

2. Consider a model m with parameters θ and hyperparameters λ. Assume the parameters θ have a prior $p(\theta|\lambda, m)$. Assume we have observed some data $\mathbf{X}$ and the likelihood is of the form $p(\mathbf{X}|\theta, \lambda, m)$. For this setup, write down the expressions for computing: (1) $p(\theta|\mathbf{X}, \lambda, m)$, (2) $p(\lambda|\mathbf{X}, m)$, and (3) $p(m|\mathbf{X})$. Rank these three quantities in terms of the hardness of computing them (easiest to hardest) and briefly justify your ranking. Note: Consider the general case; your answers should not assume any conjugacy. Also, all the quantities in your expressions should clearly and explicitly show everything you need to condition on.

3. Consider a regression model where each input $\boldsymbol{x}_n \in \mathbb{R}^D$ and output y_n is a non-negative count. Assume $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \text{Poisson}(y_n|\lambda_n)$ where $\lambda_n = \exp(\boldsymbol{w}^\top \boldsymbol{x}_n)$ and we have N training examples $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$.
(1) Derive the MLE objective function for this model. Is it possible to get MLE solution in closed form?
(2) Assume a Gaussian prior $p(\boldsymbol{w}) = \mathcal{N}(0, \sigma^2\mathbf{I})$. Is the posterior of $\boldsymbol{w}$ available in closed form. Justify your answer. Regardless of the answer, suppose we decide to use a Laplace approximation for the posterior of $\boldsymbol{w}$. Will Laplace approximation be a good idea here? Briefly justify your answer.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: September 19, 2022

4. Suppose we have N observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn i.i.d. from the exponential distribution, i.e., $p(x_n|\theta) = \theta \exp(-\theta x_n)$ and the prior on the parameter $\theta > 0$ is $p(\theta) = \text{Gamma}(\theta|a, b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}\exp(-b\theta)$. What is the marginal likelihood $p(\mathbf{X}|a, b)$? Give your answer as a closed-form expression (not an integral). Avoid very detailed derivation; show only the basic steps and write down the final expression.

5. Consider a regression model where the joint distribution of any input $\boldsymbol{x} \in \mathbb{R}^D$ and its output $y \in \mathbb{R}$ is $p(\boldsymbol{x}, y) = \frac{1}{N}\sum_{n=1}^{N} f(\boldsymbol{x} - \boldsymbol{x}_n, y - y_n)$ where $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$ denotes the training examples. Further assume $f(\boldsymbol{x} - \boldsymbol{x}_n, y - y_n) = \mathcal{N}([\boldsymbol{x} - \boldsymbol{x}_n, y - y_n]^\top|\mathbf{0}, \sigma^2\mathbf{I}_{D+1})$. For this model, derive the conditional distribution of the output y given the input, i.e., $p(y|\boldsymbol{x})$, as well as the expectation $\mathbb{E}[y|\boldsymbol{x}]$. Also give a brief justification as to why the expressions $p(y|\boldsymbol{x})$ and $\mathbb{E}[y|\boldsymbol{x}]$ make intuitive sense.

Some distributions and their properties:

- For $x \in (0,1)$, $\text{Beta}(x|a,b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$, where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and Γ denotes the gamma function s.t. $\Gamma(x) = (x-1)!$ for a positive integer x. Expectation of a Beta r.v.: $\mathbb{E}[x] = \frac{a}{a+b}$.

- For $x \in \{0,1,2,\ldots\}$ (non-negative integers), $\text{Poisson}(x|\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$ where λ is the rate parameter.

- For $x \in \mathbb{R}_+$, $\text{Gamma}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$ (shape and rate parameterization), and $\text{Gamma}(x|a,b) = \frac{1}{\Gamma(a)b^a} x^{a-1} \exp(-\frac{x}{b})$ (shape and scale parameterization)

- For $x \in \mathbb{R}$, Univariate Gaussian: $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$

- For $x \in \mathbb{R}^D$, D-dimensional Gaussian: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$.
  Trace-based representation: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}\text{trace}\left[\boldsymbol{\Sigma^{-1}}\mathbf{S}\right]\right\}$, $\mathbf{S} = (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top$.
  Information form: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-D/2}|\boldsymbol{\Lambda}|^{1/2} \exp\left[-\frac{1}{2}\left(\boldsymbol{x}^\top\boldsymbol{\Lambda}\boldsymbol{x} + \boldsymbol{\xi}^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi} - 2\boldsymbol{x}^\top\boldsymbol{\xi}\right)\right]$ where $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\xi} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$

- For $\boldsymbol{\pi} = [\pi_1,\ldots,\pi_K]$, s.t. $\sum_{k=1}^K \pi_k = 1$, $\text{Dirichlet}(\boldsymbol{\pi}|\alpha_1,\ldots,\alpha_K) = \frac{1}{B(\alpha_1,\ldots,\alpha_K)} \prod_{k=1}^K \pi_k^{\alpha_k-1}$
  where $B(\alpha_1,\ldots,\alpha_K) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^K \alpha_k)}$, and $\mathbb{E}[\pi_k] = \frac{\alpha_k}{\sum_{k=1}^K \alpha_k}$

- For $x_k \in \{0,N\}$ and $\sum_{k=1}^K x_k = N$, $\text{multinomial}(x_1,\ldots,x_K|N,\boldsymbol{\pi}) = \frac{N!}{\boldsymbol{x}_1!\ldots,x_K!}\pi_1^{x_1}\ldots\pi_K^{x_K}$
  where $\boldsymbol{\pi} = [\pi_1,\ldots,\pi_K]$, s.t. $\sum_{k=1}^K \pi_k = 1$. The multinoulli is the same as multinomial with $N=1$.

Some other useful results:

- If $\boldsymbol{x} = \mathbf{A}\boldsymbol{z} + \boldsymbol{b} + \epsilon$, $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})$, $p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0},\mathbf{L}^{-1})$ then $p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{z}+\boldsymbol{b},\mathbf{L}^{-1})$,
  $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mathbf{A}\boldsymbol{\mu}+\boldsymbol{b},\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^\top+\mathbf{L}^{-1})$, and $p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\Sigma}\left\{\mathbf{A}^\top\mathbf{L}(\boldsymbol{x}-\boldsymbol{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\right\},\boldsymbol{\Sigma})$,
  where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^\top\mathbf{L}\mathbf{A})^{-1}$.

- Marginal and conditional distributions for Gaussians: $p(\boldsymbol{x}_a) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_a,\boldsymbol{\Sigma}_{aa})$,
  $p(\boldsymbol{x}_a|\boldsymbol{x}_b) = \mathcal{N}(\boldsymbol{x}_a|\boldsymbol{\mu}_{a|b},\boldsymbol{\Sigma}_{a|b})$ where $\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa}-\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$, $\boldsymbol{\mu}_{a|b} = \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b)\right\} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b) = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b-\boldsymbol{\mu}_b)$, where symbols have their usual meaning. :)

- $\frac{\partial}{\partial\boldsymbol{\mu}}[\boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu}] = [\mathbf{A}+\mathbf{A}^\top]\boldsymbol{\mu}$, $\frac{\partial}{\partial\mathbf{A}}\log|\mathbf{A}| = \mathbf{A}^{-\top}$, $\frac{\partial}{\partial\mathbf{A}}\text{trace}[\mathbf{A}\mathbf{B}] = \mathbf{B}^\top$

- For a random variable vector $\boldsymbol{x}$, $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\top + \text{cov}[\boldsymbol{x}]$

Name:

Roll No.: Dept.: