Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: February 24, 2023

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: 5 x 2 = 10 marks). (Tick/circle all options that you think are true)

1. Which of the following is true about probabilistic linear regression with Gaussian likelihood, Gaussian prior on weights $\boldsymbol{w}$, and gamma priors on the precision hyperparametes of the likelihood and Gaussian prior? (1) The joint posterior distribution over all the unknowns can be computed in closed form, (2) The joint posterior is intractable, (3) The PPD can be computed in closed form, (4) The PPD is intractable.

2. Which of the following is true about Gaussian Processes, assuming the hyperparameters are fixed/known: (1) The PPD is available in closed form for the regression setting with Gaussian likelihood, (2) The PPD is available in closed form for settings when the likelihood is form exponential family, (3) The PPD computation, in general, does not require the posterior to be computed, (4) The label prediction for a test input only depends on the labels of a subset of training examples.

3. Which of the following is true about the marginal likelihood? (1) It is available in closed form if the likelihood and prior are a conjugate pair from exponential family, (2) It is the expectation of the likelihood w.r.t. the posterior distribution of the model parameters, (3) It is the expectation of the likelihood w.r.t. the prior distribution of the model parameters, (4) If the marginal likelihood is intractable for a model, the posterior will also be intractable.

4. Which of the following is true about MAP estimation: (1) It solution more robust against overfitting as compared to the MLE solution, (2) Assuming the log-posterior function is differentiable, computing the MAP solution is not much harder as compared to computing the MLE solution, (3) The MAP estimate is equal to the mean of the posterior, (4) When the likelihood and prior are a conjugate pair from exponential family, MAP and MLE solutions are identical.

5. Which of the following is true about the expectation maximization (EM) algorithm used for models that contain both parameters Θ as well as latent variables $\mathbf{Z}$? (1) It can be used to compute the MLE solution of the parameters, (2) It can be used to compute the MAP solution of the parameters, (3) It can be used to compute the joint posterior of the latent variables and the parameters, (4) The maximization (M) step estimates the parameters Θ by maximizing the log-likelihood $\log p(\mathbf{X}|\Theta)$.

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

1. Draw the complete plate notation diagram for a beta-Bernoulli model of N observations $y_1, y_2, \ldots, y_N$, with each $y_n \sim \text{Bernoulli}(y_n|\pi)$ and $\pi \in (0,1)$ given a $\text{Beta}(a,b)$ prior. Assume a, b to be known.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: February 24, 2023

---

2. The gradient expression for canonical GLM is of the form $\boldsymbol{g} = \sum_{n=1}^{N}(y_n - \mu_n)\boldsymbol{x}_n$, where μ_n is the conditional mean of response y_n, defined as $f(\boldsymbol{w}^\top \boldsymbol{x}_n)$. Briefly explain why this expression makes intuitive sense.

3. Briefly describe how we can construct a scale-mixture of Gaussian distributions, and also give the mathematical expression for this construction.

4. Consider the following distribution: $p(\theta|m_0, \phi_0) = \exp(\phi_0^\top \theta - m_0 g(\theta) - A(m_0, \phi_0))$. Is this an exponential family distribution? If yes, write down its natural parameters and sufficient statistics. It not, state why it is not an exponential family distribution.

5. In a latent variable model, suppose we want to perform hybrid estimation of the unknowns, i.e., compute the posterior for some and compute the point estimate for the others. How would you decide which of these approaches to use for which unknowns?

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: February 24, 2023

6. Why are the Gaussian Process based models for regression/classification slow at test time?

Section 3 (4 not-so-short answer questions: 10+8+8+6 = 32 marks). .

1. Consider a linear regression model with the likelihood $p(y_n|\boldsymbol{w}, \boldsymbol{x}_n) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta_n^{-1})$ and prior $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \boldsymbol{\Lambda}^{-1})$, where $\boldsymbol{\Lambda}$ is a diagonal precision matrix with its d^{th} diagonal entry being λ_d. Assume $\beta, \boldsymbol{\Lambda}$ to be known. The goal is to estimate $\boldsymbol{w}$ from training data $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$. Write down the final expressions for MLE and MAP objective functions (no need to solve for $\boldsymbol{w}$). Looking at these expressions, what roles do β_n and λ_d play here? Also write down the final expression for the posterior distribution of $\boldsymbol{w}$. You need not show the full derivations; only the final expressions are required.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: February 24, 2023

2. Consider a generative classification model with training data $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$ where $\boldsymbol{x}_n \in \mathbb{R}^D$ and $y_n \in \{0, 1\}$. Assume each class-conditional distribution to be a Gaussian with its covariance matrix being known. Which quantities would you need to estimate for training this model? Derive the expressions for the MLE solutions of these quantities. Please do not show very detailed steps of derivations; only write the key equations and the final answers. Would the expressions for MAP solutions of these quantities, in general, be the same as the MLE solution? Briefly justify the answer.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: February 24, 2023

3. Consider a logistic regression model $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \frac{1}{1+\exp(-y_n \boldsymbol{w}^\top \boldsymbol{x}_n)}$, with a zero-mean Gaussian prior $p(\boldsymbol{w}) = \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I})$. Note that this loss function for logistic regression assumes $y_n \in \{-1, +1\}$ instead of $\{0, 1\}$. Show that the MAP estimate for $\boldsymbol{w}$ can be written as $\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y_n \boldsymbol{x}_n$ where each α_n itself is a function of $\boldsymbol{w}$. Based on the expression of α_n, you would see that it has a precise meaning. Briefly state what α_n means, and also briefly explain why the result $\boldsymbol{w} = \sum_{n=1}^{N} \alpha_n y_n \boldsymbol{x}_n$ makes sense for this model.

4. Consider N scalar-valued observations $x_1, \ldots, x_N$ drawn i.i.d. from $\mathcal{N}(\mu, \sigma^2)$. Consider their empirical mean $\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n$. Expressing $\bar{x}$ as a linear transformation of a random variable, derive the probability distribution of $\bar{x}$.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: February 24, 2023

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: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\Big\{ -\frac{1}{2}\text{trace}\,[\boldsymbol{\Sigma}^{-1}\mathbf{S}] \Big\}$, $\mathbf{S} = (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top$.
  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.**:

FOR ROUGH WORK ONLY

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Mid-sem Exam
Date: February 24, 2023