Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: November 21, 2022

Instructions: *Total:* **100 marks**

1. Total duration: **3 hours**. Please write your name, roll number, department on **all pages**.
2. This booklet has 10 pages (9 pages + 1 page 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 (True or False: 15 X 1 = 15 marks). For each of the following simply write **T** or **F** in the box.

1. [ ] The frequentist approach cannot provide an estimate of the parameter uncertainty.
2. [ ] SGLD sampling based inference usually converges faster than Metropolis-Hastings.
3. [ ] In Metropolis-Hastings sampling for posterior inference, the cost of calculating the acceptance probability depends on the dataset size.
4. [ ] The predictive variance of a regression model when using the MAP estimate of the weights depends on the inputs.
5. [ ] When using VI to approximate the posterior, the posterior predictive distribution does not require an approximation.
6. [ ] A non-probabilistic autoencoder cannot be used to generate new data.
7. [ ] A standard generative adversarial network (GAN) cannot be used to compress/encode data.
8. [ ] If a node c is the common parent of two nodes a and b and there are no other edges between nodes, then conditioned on c, nodes a and b are independent.
9. [ ] A generative model for classification, in general, has fewer parameters to estimate than a discriminative model.
10. [ ] A generalized linear model (GLM) is guaranteed to have a closed form solution for its parameters when doing MLE/MAP estimation.
11. [ ] Storing the VI approximation of a posterior distribution requires less space as compared to its MCMC approximation.
12. [ ] When using the Bayes rule to find the expression for the posterior, the denominator in the Bayes rule expression (i.e., the marginal likelihood) can always be ignored.
13. [ ] Active Learning is based on querying the labels of those unlabeled examples whose predicted label distribution has the smallest entropy.
14. [ ] The posterior predictive distribution of Bayesian linear regression with Gaussian likelihood, Gaussian prior on weights, and rest of the hyperparameters being unknown, is not tractable.
15. [ ] A large value of the concentration parameter α in a Dirichlet Process based nonparametric Bayesian model for clustering will lead to large inferred number of clusters.

Section 2 (15 short answer questions: 15 x 3 = 45 marks). .

1. How is Bayesian ML different from "conventional" ML? Specifically, state the difference in terms of: (a) How parameter estimation is done? (b) How predictions are made?

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: November 21, 2022

2. Given two models, $\mathcal{M}_1$ and $\mathcal{M}_2$ with their average accuracies and average confidence scores being a_1, c_1 and a_2, c_2, respectively, in a rough sense how would be find out which model has a better calibration?

3. Suppose you have learned the latent Dirichlet allocation (LDA) model on some text data. Given a topic vector ϕ_k, how would you find out what topic (in terms of its semantic meaning) ϕ_k represents?

4. Briefly explain how a shrinkage based approach to nonparametric Bayesian modeling helps learn the right model size. You may use an example of a model discussed in the class.

5. Why would a method like Rejection Sampling not be an ideal choice for doing posterior inference for a model like Bayesian linear regression or logistic regression or various other Bayesian ML models?

6. Laplace approximation can be expensive for Bayesian neural networks due to the enormous number of parameters. How is the Laplace approximation used in a hybrid Bayesian neural network?

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: November 21, 2022

7. In what way is the deep ensemble approach better than inference methods such as sampling or VI in terms of the quality of inferred posterior?

8. (Bayesian) Active Learning and Bayesian Optimization use an acquisition function to decide which observation(s) to acquire next. Briefly describe the basic difference between the goals of the respective acquisition functions used in both these methods?

9. Can you combine a standard (not Bayesian) deep neural network and Bayesian linear regression model to do Bayesian *nonlinear* regression? If yes, how? If no, why not?

10. Briefly explain why VI which is based on minimizing $\text{KL}[q(Z)||p(Z|X)] = \int q(Z) \log \frac{q(Z)}{p(Z|X)} dZ$ tends to underestimate the variance of the true posterior $p(Z|X)$.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: November 21, 2022

11. State 3 advantages (should be distinct from each-other) of Gaussian Process based models over standard, kernelized supervised learning models, such as SVM, or methods such as nearest neighbors.

12. Consider a likelihood model $p(x|\theta) = \mathcal{N}(x|\theta, \sigma^2)$ with σ^2 known, and assume a prior distribution $p(\theta) = \exp(A\theta^2 + B\theta + C)$. Is $p(\theta)$ conjugate to the likelihood $p(x|\theta)$? If yes, why? If no, why not?

13. Suppose you have a coin with bias (probability of a head) π having a prior $\text{Beta}(\pi|a, b)$. You toss the coin until you see a total of k heads (assume i.i.d. tosses). Note that the number of heads is fixed here but the total number (call it n) of tosses is stochastic. Give the expression for the likelihood, and also the posterior of π (no need to derive it; just the final answer is fine if you can find it by inspection/intuition).

14. What is the benefit of using amortized inference for inferring the local latent variables?

15. What is the meaning of "collapsing" in context of Bayesian inference? What is the benefit of collapsing?

Name:

Roll No.: **Dept.:**

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: November 21, 2022

Section 3 (4 not-so-short answer questions: 10+8+12+10 = 40 marks). .

1. Suppose you are given N data-points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ where each data-point $\boldsymbol{x}_n \in \mathbb{R}^D$ has some missing features. Denote it as $\boldsymbol{x}_n = [\boldsymbol{x}_n^{(o)}, \boldsymbol{x}_n^{(m)}]$, where $\boldsymbol{x}_n^{(o)}$ denotes the observed features and $\boldsymbol{x}_n^{(m)}$ denotes the missing features. The set of features missing could be different in different data-points. Assume each data-point $\boldsymbol{x}_n$ is generated from a Gaussian $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Give an EM algorithm to compute the MLE of the parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ of this Gaussian (in the box below, you should only mention the key steps/equations in the EM algorithm and skip detailed derivations) .
   Hint: Treat the missing features $\boldsymbol{x}_n^{(m)}$ of each data-point as latent variables.
   Note: You may use the standard result that the MLE of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a Gaussian when there are no missing features is given by $\boldsymbol{\mu}_{MLE} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n$, and $\boldsymbol{\Sigma}_{MLE} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \boldsymbol{\mu}_{MLE})(\boldsymbol{x}_n - \boldsymbol{\mu}_{MLE})^\top$

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: November 21, 2022

2. Consider a K component Gaussian mixture model with parameters $\Theta = \{\pi_k, \boldsymbol{\mu}_k, \sigma_k^2\mathbf{I}\}_{k=1}^K$. Note that the covariance matrices of all the Gaussians are assumed to be spherical. Consider the posterior distribution $p(\boldsymbol{z}|\boldsymbol{x}, \Theta)$, of the cluster assignment $\boldsymbol{z} \in \{1, \ldots, K\}$ for an observation $\boldsymbol{x} \in \mathbb{R}^D$.

   Write down the expression for $p(\boldsymbol{z}|\boldsymbol{x}, \Theta)$. What will be the entropy of this posterior distribution if the variances σ_k^2 of each Gaussian are set to a very small value (close to zero)? Justify your answer. Also, a particular type of clustering algorithm arises in this special case. What is that algorithm?

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: November 21, 2022

3. Consider modeling an $N \times K$ binary matrix $\mathbf{Z}$ with a prior such that its binary entries are assumed to be generated i.i.d. as follows

$$\begin{aligned} \pi_k &\sim \text{Beta}(\alpha/K, 1) & k = 1, \ldots, K \\ Z_{nk}|\pi_k &\sim \text{Bernoulli}(\pi_k) & n = 1, \ldots, N, k = 1, \ldots, K \end{aligned}$$

- Derive the expression for the marginal prior $p(\mathbf{Z}|\alpha)$ after integrating out the π_k's.
- Derive the expression for $p(Z_{nk}|Z_{-nk})$ where Z_{-nk} denotes all the entries of $\mathbf{Z}$, except Z_{nk}. What will $p(Z_{nk}|Z_{-nk})$ be as $K \to \infty$? What does the result mean intuitively?
- (Bonus: 5 marks) As a function of α, what will be the expected number of ones in each column of $\mathbf{Z}$, and in all of $\mathbf{Z}$?

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: November 21, 2022

4. Consider a regression problem where we are modeling count-valued responses using a Poisson distribution $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \text{Poisson}(y_n|\exp(\boldsymbol{w}^\top \boldsymbol{x}_n))$, $n = 1, \ldots, N$. Assume a Gaussian prior on the weights, i.e., $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|0, \lambda^{-1}\mathbf{I})$ with λ known. The goal is to infer the posterior of $\boldsymbol{w}$. Given two choices for the inference algorithm, MH sampling and SGLD, which one would you prefer for this problem and why? Give a sketch of your algorithm of choice and give the necessary update equations.

Name:

Roll No.: **Dept.:**

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: November 21, 2022

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.**: