Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: April 29, 2023

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. [ ] Gibbs sampling is applicable even if there is no (local) conjugacy.
2. [ ] Choosing the likelihood and the prior as exponential family distributions results in a closed form posterior distribution.
3. [ ] Unlike MCMC, variational inference gives point estimates of the unknowns.
4. [ ] Denoising diffusion models are slower at generation as compared to GAN or VAE
5. [ ] In active learning, inputs for which the current model is least confident in its prediction are likely to be more informative.
6. [ ] EM does not provide any uncertainty estimates for any of the unknowns of the model.
7. [ ] The true joint posterior of all the unknowns of a model is equal to the product of their conditional posteriors.
8. [ ] EM algorithm yields local optima of local variables and global optima for global variables.
9. [ ] A symmetric proposal distribution makes the calculation of acceptance probability significantly faster in MCMC.
10. [ ] Monte-Carlo dropout based approximation of the posterior can be obtained from the point estimates of the model parameters.
11. [ ] SWAG and Laplace approximation both yield the same approximation of the posterior.
12. [ ] Likelihood is a function, not a probability distribution.
13. [ ] It is not possible to do MLE or MAP to estimate the parameters of a generative adversarial network (GAN).
14. [ ] Monte Carlo sampling can be used to compute both ELBO as well as its derivatives.
15. [ ] Generalized linear model (GLM) is a generative model.

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

1. If the target distribution has multiple modes, standard SGLD is prone to generating samples near one of the modes. How can SLGD, or other gradient based sampling methods similar to SGLD, address this issue and generate samples from around the multiple modes of the distribution? Justify your answer briefly.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: April 29, 2023

2. Give two reasons as to why Gaussian Process (GP) is a good method to estimate the surrogate model of the function being optimized via Bayesian Optimization.Why won't you use a Bayesian linear regression model for this purpose?

3. Briefly explain how entropy of the posterior distribution of model parameters can be used for active learning.

4. Suppose you have run MCMC to generate (a sufficiently large number of) samples from a distribution $p(\boldsymbol{z})$. How would you use the generated samples to find the maxima (mode) of this distribution (need not be the true mode; an approximate mode is fine)?

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: April 29, 2023

5. Can the integral $\int_{-\infty}^{\infty} \exp[-\lambda(x-\mu)^2]dx$ be computed exactly? If yes, write its value. If no, state why.

6. Suppose you have tossed a coin a number of times. Now suppose you want to compute the probability that $\theta \leq 0.4$ where θ is the probability of heads. Briefly suggest a Bayesian way to do this.

7. Which of these inference algorithms can be used to infer the posterior over the weights $\boldsymbol{w}$ of logistic regression model, assuming no additional variables are introduced for the model: (1) Expectation-Maximization, (2) Gibbs Sampling, (3) Metropolis-Hastings Sampling, (4) Stochastic Gradient Langevin Dynamics? Briefly justify your answers.

8. Can we use the generative approach to learn a regression model? If yes, can it be done in the same way as we learn a classification model using a generative approach, i.e., defining $p(y|\boldsymbol{x}) = \frac{p(y)p(\boldsymbol{x}|y)}{p(\boldsymbol{x})}$? Briefly justify your answer.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: April 29, 2023

9. Consider a model with data X from a likelihood model $p(X|\theta, \beta)$ and prior $p(\theta|\alpha)$ on the parameters θ. Briefly state how the two hyperparameters β, α of this model (one is part of likelihood and the other is part of the prior) can be estimated using expectation-maximization (EM) if (1) We want their point estimates, and (2) If we want their conditional posterior distributions.

10. 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.

11. What is the difference between variational inference and variational EM? When would you need to use variational EM as opposed to standard EM?

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: April 29, 2023

12. What is the advantage of selecting the "best" hyperparameter values using an MLE-II approach as compared to using cross-validation?

13. A zero-mean Gaussian prior is equivalent to using L2 regularization on the weight vector w. Can such a prior be used to impose different amounts of regularization on different components of the weight vector? If yes, how? If no, why not?

14. Briefly state why the marginal likelihood of a model can also be seen as a special case of the posterior predictive distribution.

15. Assume you have K candidate models (assume probabilistic models) that you can possibly try out for a classification problem and don't know which one is the "best". Briefly explain how would a fully Bayesian approach handle this problem.

Name:

Roll No.: **Dept.:**

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: April 29, 2023

Section 3 (4 not-so-short answer questions: 8+8+8+16 = 40 marks). .

1. Assume N observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn i.i.d. from the exponential distribution,which is defined as $p(x_n|\theta) = \theta \exp(-\theta x_n)$ 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.

2. Briefly describe what collapsing means in the context of approximate Bayesian inference, and what are the benefits of collapsing? You may use an example, such as a model like Gaussian mixture model or Latent Dirichlet Allocation. You don't need to be excessively detailed (i.e., no derivations etc); it would suffice to explain via a simple example the basic difference between the uncollapsed vs collapsed inference.

Name:

Roll No.: **Dept.:**

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: April 29, 2023

3. Suppose you are given a dataset of N labeled images $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$, where each image is either a picture of a cat or a dog. You want to train a logistic regression model with weights $\boldsymbol{w}$ to predict whether a new image is of a cat or a dog. However, some of the images in the dataset are mislabeled (i.e., some of the images labeled as "cat" are actually dogs, and vice versa). Let's assume that the mislabeling is random, such that each image is mislabeled with probability p, independently of all other images.

   (a) Write the expression for the likelihood function for this problem. (3 marks)

   (b) Assuming a suitable prior that corresponds to an $L2$ regularizer, write the expression for the posterior distribution of $\boldsymbol{w}$. You only need to write it up to a proportionality constant. (2 marks)

   (c) Given the posterior distribution, how would you compute the probability of correctly classifying a new image? How does this probability depend on the mislabeling probability p? (3 marks)

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: April 29, 2023

4. Consider a linear regression model $\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}$ with $\boldsymbol{y} = [y_1, \ldots, y_N]^\top$ is the $N \times 1$ response vector, $\mathbf{X}$ is the $N \times D$ feature matrix, and $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^\top$ is the $N \times 1$ vector of i.i.d. Gaussian noise $\mathcal{N}(0, \sigma^2)$. Let us assume the following prior on each entry of the weight vector $\boldsymbol{w} \in \mathbb{R}^D$

$$p(w_d|\sigma, \gamma_d) = \begin{cases} \mathcal{N}(0, \sigma^2 v_0), & \text{if } \gamma_d = 0 \\ \mathcal{N}(0, \sigma^2 v_1), & \text{if } \gamma_d = 1 \end{cases}$$

where $v_1 \gg v_0 > 0$. Further assume the priors $p(\gamma_d) = \text{Bernoulli}(\theta)$, $d = 1, \ldots, D$, $p(\theta) = \text{Beta}(a_0, b_0)$, and $p(\sigma^2) = \text{IG}(\nu/2, \nu\lambda/2)$, where IG denotes the inverse-gamma prior in its shape-scale paramerization. Note that theprior on w_d can also be written as $p(w_d|\sigma, \gamma_d) = \mathcal{N}(0, \sigma^2 \kappa_{\gamma_d})$ with $\kappa_{\gamma_d} = \gamma_d v_1 + (1 - \gamma_d) v_0$.

- What is the effect of assuming the above prior on $\boldsymbol{w}$ (4 marks)?
- Derive an EM algorithm for this model. Your algorithm should give the posterior over the weight vector $\boldsymbol{w}$ and point estimates (MAP) for the remaining unknowns $\boldsymbol{\gamma}, \sigma^2, \theta$ (12 marks).

(continue on next page)

Name:

Roll No.:

Dept.:

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
End-sem Exam
Date: April 29, 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: $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2}\text{trace}\left[\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\right\}$, $\mathbf{S} = (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top$.
  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 $\boldsymbol{x}$, $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{x}]^\top + \text{cov}[\boldsymbol{x}]$