Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Quiz-3
Date: March 18, 2024

Instructions:

Total: **25 marks**

1. Please write your name, roll number, department on **all pages** of this question paper.
2. Write your answers clearly in the provided box. Keep your answer precise and concise.

Section 1 (10 short answer questions: 3+2+4+3+2+2+2+2+2+3 = 25 marks). .

1. Consider a Bayesian linear regression model with unknowns $\boldsymbol{w} \in \mathbb{R}^D, \lambda \in \mathbb{R}_+, \beta \in \mathbb{R}_+$. Suggest a reasonable mean-field VI approximation of the variational distribution $q(\boldsymbol{w}, \lambda, \beta)$ which approximates the joint posterior. What is the total number of variational parameters in your mean-field VI approximation?

   We can use $q(\boldsymbol{w}, \lambda, \beta) = q(\boldsymbol{w})q(\lambda)q(\beta)$ where $q(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\mu, \Sigma)$, and $q(\lambda)$ and $q(\beta)$ can be gamma distributions. Assuming Σ to be diagonal, the total number of variational parameters would be $2D + 4$ (basically D parameters for μ, D for Σ, and 2 parameters each for each of the gamma distributions).

2. Suppose we have two models $\mathcal{M}_1$ and $\mathcal{M}_2$ and we use VI to approximate the posterior for the unknowns of each model. Can VI also help you decide which model is better? If yes, why? If no, why not?

   We can compare the ELBO since it is a lower bound on the log-marginal likelihood and pick the model with the larger value of ELBO.

3. Consider a model with unknowns that consist of latent variables $\mathbf{Z}$ and parameters Θ. Given data $\mathcal{D}$ and assuming a variational distribution $q_\phi(\mathbf{Z})$, the ELBO is $\mathcal{L}(\phi, \Theta) = \mathbb{E}_{q_\phi(\mathbf{Z})}[\log p(\mathcal{D}, \mathbf{Z}|\Theta) - \log q_\phi(\mathbf{Z})]$. Assuming intractable ELBO, briefly describe how to compute a Monte-Carlo approximation of ELBO's gradient w.r.t. parameters Θ, and write the general expression of this Monte-Carlo approximation. Can the ELBO's gradient w.r.t. ϕ be computed in as identical fashion as Θ? Briefly justify your answer.

   ELBO's gradient w.r.t. Θ is $\nabla_\Theta \mathcal{L}(\phi, \Theta) = \nabla_\Theta \mathbb{E}_{q_\phi(\mathbf{Z})}[\log p(\mathcal{D}, \mathbf{Z}|\Theta) - \log q_\phi(\mathbf{Z})] = \mathbb{E}_{q_\phi(\mathbf{Z})} \nabla_\Theta [\log p(\mathcal{D}, \mathbf{Z}|\Theta) - \log q_\phi(\mathbf{Z})]$ for which we can easily get a Monte-Carlo approximation given M samples $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(M)}$ from $q_\phi(\mathbf{Z})$ and thus $\nabla_\Theta \mathcal{L}(\phi, \Theta) \approx \frac{1}{M} \sum_{i=1}^M [\nabla_\Theta \log p(\mathcal{D}, \mathbf{Z}^{(i)}|\Theta) - \nabla_\Theta \log q_\phi(\mathbf{Z}^{(i)})]$. We can't do the same when computing ELBO's gradient w.r.t. ϕ because the order of ∇_ϕ and expectation w.r.t. $q_\phi(\mathbf{Z})$ can't be interchanged in the same manner, and we need to use methods like BBVI or reparametrization trick.

4. Rejection sampling generates samples from a distribution $p(z) = \frac{\tilde{p}(z)}{Z_p}$ using a proposal distribution $q(z)$ such that $Mq(z) \geq \tilde{p}(z), \forall z$ where $M > 0$ is a constant. Very large values of M will more easily satisfy this condition. However, what might be a disadvantage of using very large value of M?

   Very large values of M will also result in large number of samples being rejected. Think of the figure of rejection sampling illustration in which $Mq(z)$ envelops the true distribution from the above. The "gap" region will be very large, and thus a large number of samples will be rejected

5. When would you need to use importance sampling instead of standard Monte Carlo sampling to approximate an intractable expectation of the form $\mathbb{E}_p[f] = \int f(z)p(z)dz$? Briefly explain how to compute this approximation and write down its expression assuming we are using M random samples of z.

   If we can't directly sample from $p(z)$, we may use importance sampling where we instead sample from another simpler proposal distribution $q(z)$ and use these samples to approximate the expectation as $\mathbb{E}_p[f] = \int f(z)p(z)dz = \int f(z)\frac{p(z)}{q(z)}q(z)dz \approx \frac{1}{M}\sum_{i=1}^M f(z^{(i)})\frac{p(z^{(i)})}{q(z^{(i)})}$ where $\{z^{(i)}\}_{i=1}^M$ denotes M i.i.d. samples generated from the proposal distribution $q(z)$.

6. What is the basic difference between the proposal distribution used in MCMC and the proposal distribution used in other sampling methods like rejection sampling?

   MCMC sampling follows a Markov chain such that the proposal distribution depend on the previously generate sample, whereas for other sampling methods such as rejection sampling, the proposal distribution does not depend on the previous sample.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Quiz-3
Date: March 18, 2024

7. What is the advantage of methods like Langevin dynamics based sampling and Hamiltonian Monte Carlo sampling as compared to standard Metropolis-Hastings based MCMC sampling?

   Sampling methods like Langevin dynamics and HMC make use of the posterior distribution's gradient information in the sampling process which results in more efficient exploration of the target distribution. In contrast, standard MH based MCMC sampling generates samples using a random walk-like behavior (proposals don't take into account the posterior's gradient info), and are likely to be less efficient.

8. Can Gibbs sampling be applied if some of the CPs are distributions from which we can't sample directly? If yes, how? If no, why not?

   Note that each iteration of Gibbs sampling consists of multiple steps and each step itself involves sampling from a CP. If we can't sample directly from some CPs in these steps, Gibbs sampling is still applicable - we can use methods like rejection sampling or MH sampling to draw from such CPs.

9. In terms of storage requirements of VI and sampling, which of the two is more expensive and why?

   Sampling is more expensive because to represent the approximate target distribution, we need to store a large number of samples $\{Z^{(i)}\}_{i=1}^{M}$ generated by the sampling algorithm. In contrast, VI learns an explicit parametric approximation $q(Z|\phi)$ of the target distribution, which only requires storing the variational parameters ϕ.

10. Write down the basic mathematical expression for the model improvement based acquisition function $A(x_*)$ when selecting an unlabeled example x_* in active learning. Why does such an $A(x_*)$ make sense?

    Model improvement based acquisition function can be written as the difference in the entropy of the current model and the expected entropy of the updated model after including x_* in the training set. In equations, we can write it as $A(x_*) = \mathbb{H}[p(\theta|\mathcal{D})] - \int_{y_*|x_*,\mathcal{D}} \mathbb{H}[p(\theta|\mathcal{D} \cup (x_*, y_*))]dy_*$, where $\mathbb{H}$ denotes the entropy function. This acquisition function $A(x_*)$ makes sense because we are choosing an input that is helping reduce our uncertainty about the model maximally.