Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Quiz-2
Date: April 10, 2023

Instructions:

Total: **30 marks**

1 Write your answers clearly in the provided box. Keep your answer precise and concise.

Section 1 (10 very short answer questions: 10 x 3 = 30 marks). .

1. VI minimizes the KL divergence $\mathrm{KL}(q||p) = \int q(z) \log \frac{q(z)}{p(z)} dz$ between the true distribution $p(z)$ and its variational approximation $p(z)$. Using this definition of VI, briefly explain why the variational approximation $q(z)$ tends to underestimate the variance of the true distribution $p(z)$.

   The variational approximation $q(z)$ will avoid going in regions where $p(z)$ is low, otherwise KL will blow up. Therefore, $q(z)$ isn't spread as much as $p(z)$ and thus variance will be underestimated.

2. For a latent variable model with data $\mathbf{X}$, latent variables $\mathbf{Z}$, and parameters Θ, write down the objective function for the MLE problem for Θ, and the alternate objective function that EM maximizes to find the MLE for Θ. Do they both yield the same solution for Θ? Briefly justify your answer.

   MLE objective is $\log p(\mathbf{X}|\Theta)$ and the objective function maximized by EM is the expected completed data log-likelihood $\mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$ where the expectation is w.r.t. the CP of $\mathbf{Z}$, i.e., $p(\mathbf{Z}|\Theta, \mathbf{X})$. Since the expected CLL is not equal to the log-likelihood, the solutions given by both approaches won't necessarily be the same either.

3. Consider a linear model where the d-th entry for the weight vector is $w_d \in \mathbb{R}$. Let's assume w_d has a prior $\text{Laplace}(w_d|0, 1/\gamma) = \int \mathcal{N}(w_d|0, \tau_d^2)\text{Gamma}(\tau_d^2|1, \gamma^2/2)d\tau_d^2$. Will the conditional posterior for τ_d have a closed form? If yes, write its expression. If no, briefly state the reason.

   Given w_d, the CP of τ_d^2 will be proportional to the product of a Gaussian and a gamma distribution. However, here τ_d^2 is variance, not precision, so the CP of τ_d^2 won't actually be available in closed form (you may note that $\mathcal{N}(w_d|0, \tau_d^2)$ and $\text{Gamma}(\tau_d^2|1, \gamma^2/2$ don't have the same functional form in τ_d^2 so this pair isn't actually conjugate). As a side-note, if you want to compute the CP of $1/\tau_d^2$, this actually has a closed form given by "inverse Gaussian" distribution.

4. Stochastic gradient Langevin dynamics (SGLD) and Hamiltonian Monte Carlo (HMC), both, use gradient information of the posterior. What additional advantage does HMC have which makes it converge faster than SGLD to the target distribution?

   HMC also uses the momentum as an auxiliary variable which makes it more efficient in terms of the convergence to the target distribution.

5. Briefly explain how amortized variational inference achieves efficiency when used for inference in probabilistic models with local latent variables?

   Instead of estimating the parameters ϕ_n for the variational distribution $q(\boldsymbol{z}_n|\phi_n)$ of each latent variable $\boldsymbol{z}_n$, amortized VI learns a single deep neural network which takes as input $\boldsymbol{x}_n$ and outputs ϕ_n, i.e., $\phi_n = \text{NN}(\boldsymbol{x}_n, \mathbf{W})$ where $\mathbf{W}$ denotes the weights of the deep neutal network.

6. Consider a Bayesian linear regression model with likelihood $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \beta) = \mathcal{N}(y_n|\boldsymbol{w}^\top \boldsymbol{x}_n, \beta^{-1})$ and prior $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|0, \lambda^{-1}\mathbf{I})$. Briefly explain how you can use EM to estimate not just the weight vector $\boldsymbol{w}$ but also the hyperparameters β, λ of the likelihood and the prior.

   We can treat $\boldsymbol{w}$ as latent variable and β, λ as parameters, and estimate them via EM (CP for $\boldsymbol{w}$ and point estimates for β, λ). Alternatively, we can also treat β, λ as latent variables (assuming a reasonable prior on these) and $\boldsymbol{w}$ as parameters and again use EM to estimate these. Both are valid approaches.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS772A (PML)
Quiz-2
Date: April 10, 2023

---

7. When the posterior distribution is approximated using VI, is it possible in any situation to get the posterior predictive distribution (PPD) as a closed form expression? Also answer the same question in case we use MCMC to approximate the posterior.

   If the variational distribution $q(\theta|\mathbf{X})$ is such that the PPD integral $\int p(x_*|\theta)p(\theta|\mathbf{X})d\theta$ is tractable then the PPD is available in closed form. Otherwise, we will need to sample from $q(\theta|\mathbf{X})$ and use a Monte-Carlo averaging. On the other hand, in MCMC, we do not get a parametric form of the target distribution but only samples from it. Therefore, we can't get a closed-form expression for the PPD in this case and we must use Monte-Carlo averaging using the MCMC samples.

8. Briefly mention at least two situations in which sampling methods can be helpful in an EM algorithm.

   (1) For approximating the CP $p(\mathbf{Z}|\mathbf{X}, \Theta)$ of latent variables if the CP is intractable; and (2) For approximating the expected CLL $\mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\Theta)]$ if this expectation is intractable.

9. The expression for ELBO can be written as $\mathcal{L}(q) = \mathbb{E}_q[\log p(\mathbf{X}|\mathbf{Z})] - \mathrm{KL}[q(\mathbf{Z})||p(\mathbf{Z})]$. Briefly explain why the optimal q obtained by maximizing the ELBO has the desirable properties that we want.

   Maximizing the ELBO will give a variational posterior $q(\mathbf{Z}|\phi)$ such that latent variables $\mathbf{Z}$ from this distribution explain the data $\mathbf{X}$ well, i.e.., expectated log-likelihood is high (the first term on RHS), and $q(\mathbf{Z}|\phi)$ is also close to the prior $p(\mathbf{Z})$ (minimizing the KL between these two) thereby achieving a regularization like effect.

10. Consider an MCMC algorithm for sampling from a target distribution $p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z_p}$ using a proposal $\mathcal{N}(\boldsymbol{z}|\boldsymbol{z}^{(\tau)}, \sigma^2\mathbf{I})$ which generates the next state $\boldsymbol{z}^*$ given the current state $\boldsymbol{z}^{(\tau)}$. To decide whether to accept $\boldsymbol{z}^*$ as the next state, suppose we use an acceptance probability defined as $A = \min\{1, \frac{\tilde{p}(\boldsymbol{z}^{(\tau)})}{\tilde{p}(\boldsymbol{z}^*)}\}$. Does this acceptance probability make sense? If yes, why? If no, why not?

    No, this acceptance probability will have the opposite effect of what we want. We want the acceptance probability to be large if $\tilde{p}(z^*)$ is large, so this quantity should be in the numerator (as in the definition of MH/Metropolis sampling acceptance probability).