Probabilistic Machine Learning
(CS772A, Spring 2023)
Homework 3
Due Date: April 30, 2023 (11:59pm)

Instructions:

- Only electronic submissions will be accepted. Your main PDF writeup must be typeset in LaTeX (please also refer to the "Additional Instructions" below).
- Your submission will have two parts: The main PDF writeup (to be submitted via Gradescope `https://www.gradescope.com/`) and the code for the programming part (to be submitted via this Dropbox link: `https://tinyurl.com/cs772sp23hw3`). Both parts must be submitted by the deadline to receive full credit (**delay in submitting either part would incur late penalty for both parts**). We will be accepting late submissions upto 72 hours after the deadline (with every 24 hours delay incurring a 10% late penalty, applied on per-hour delay basis). We won't be able to accept submissions after that.
- We have created your Gradescope account (you should have received the notification). Please use your IITK CC ID (not any other email ID) to login. Use the "Forgot Password" option to set your password.

Additional Instructions

- We have provided a LaTeX template file `hw3sol.tex` to help typeset your PDF writeup. There is also a style file `pmi.sty` that contain shortcuts to many of the useful LaTeX commends for doing things such as boldfaced/calligraphic fonts for letters, various mathematical/greek symbols, etc., and others. Use of these shortcuts is recommended (but not necessary).
- Your answer to every question should begin on a new page. The provided template is designed to do this automatically. However, if it fails to do so, use the `\clearpage` option in LaTeX before starting the answer to a new question, to *enforce* this.
- While submitting your assignment on the Gradescope website, you will have to specify on which page(s) is question 1 answered, on which page(s) is question 2 answered etc. To do this properly, first ensure that the answer to each question starts on a different page.
- Be careful to flush all your floats (figures, tables) corresponding to question n before starting the answer to question $n+1$ otherwise, while grading, we might miss your important parts of your answers.
- Your solutions must appear in proper order in the PDF file i.e. solution to question n must be complete in the PDF file (including all plots, tables, proofs etc) before you present a solution to question $n+1$.
- For the programming part, all the code and README should be zipped together and submitted as a single file named `yourrollnumber.zip`. Please DO NOT submit the data provided.

Problem 1 (20 marks)

Standard VI minimizes the KL divergence between the true distribution $p(z)$ and itse variational approximation $q(z)$. A more general form of divergence is the α-divergence defined as

$$D_\alpha(p||q) = \frac{4}{1-\alpha^2}\left(1 - \int p(z)^{(1+\alpha)/2} q(z)^{(1-\alpha)/2} dz\right)$$

Show that $KL(p||q)$ corresponds to α-divergence as $\alpha \to 1$. You may use the result $p^\epsilon = \exp(\epsilon \log p) = 1 + \epsilon \log p + O(\epsilon^2)$ and then take the limit $\epsilon \to 0$. Likewise, show that $KL(q||p)$ corresponds to $\alpha \to -1$.

Problem 2 (20 marks)

Consider N scalar-valued observations $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ assumed generated i.i.d. from a Gaussian $\mathcal{N}(\mu, \tau^{-1})$ with its parameters having priors $p(\mu) = 1/\sigma_\mu$ and $p(\tau) = 1/\tau$. Assume σ_μ to be a constant. Your goal is to use mean-field variational inference to approximate the true posterior $p(\mu, \tau|\mathbf{X})$ by a variational distribution $q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau)$.

Using the mean-field VI recipe, i.e., each optimal factor in the mean-field approximation satisfies the identity $\log q^*(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\log p(\mathbf{X}, \mathbf{Z})]$ + const, find out what the distributions $q_\mu(\mu)$ and $q_\tau(\tau)$ are and clearly write down the optimal parameters of these distributions. In your PDF writeup, please only show the key steps of the derivations.

Problem 3 (40 marks)

Consider the Latent Dirichlet Allocation (LDA) model

$$\begin{aligned}
\phi_k &\sim \text{Dirichlet}(\eta, \ldots, \eta), \quad k = 1, \ldots, K \\
\theta_d &\sim \text{Dirichlet}(\alpha, \ldots, \alpha), \quad d = 1, \ldots, D \\
\boldsymbol{z}_{d,n} &\sim \text{multinoulli}(\theta_d), \quad n = 1, \ldots, N_d \\
\boldsymbol{w}_{d,n} &\sim \text{multinoulli}(\phi_{\boldsymbol{z}_{d,n}})
\end{aligned}$$

In the above, ϕ_k denotes the V dim. topic vector for topic k (assuming vocabulary of V unique words), θ_d denotes the K dim. topic proportion vector for document d, and the number of words in document d is N_d.

Your task is to derive a Gibbs sampler for the word-topic assignment variable $z_{d,n}$ (for each word in each document). Your sampler should not sample β_k, θ_d but only be sampling the $z_{d,n}$'s from the conditional posterior (CP). So you need to derive a *collapsed Gibbs sampler* by integrating out β_k and θ_d from the prior/likelihood (whichever needs to be integrating out) when computing the desired CPs.

Derive and clearly write down the the expressions for the CP that the Gibbs sampler requires in this case, and sketch the overall Gibbs sampler. Important: None of the expressions should contain θ_d and ϕ_k. Also briefly justify why your expression for CP makes intuitive sense.

Suppose, in addition, we are also interested in computing the posterior expectation $\mathbb{E}[\theta_d]$ for each document and the posterior expectation $\mathbb{E}[\phi_k]$ for each topic, using the information in the collected samples of $\mathbf{Z}$. Suggest a way and give the proper expressions (approximation is fine) that compute these quantities, and give an intuitive meaning of the final expressions for $\mathbb{E}[\theta_d]$ and $\mathbb{E}[\phi_k]$.

Problem 4 (20 marks)

Consider a matrix factorization model for a partially observed $N \times M$ matrix $\mathbf{R}$, where $p(r_{ij}|\boldsymbol{u}_i, \boldsymbol{v}_j) = \mathcal{N}(r_{ij}|\boldsymbol{u}_i^\top \boldsymbol{v}_j, \beta^{-1})$, and $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ denote the latent factors of i-th row and j-th column or $\mathbf{R}$, respectively. The posterior predictive distribution of each r_{ij} is defined as $p(r_{ij}|\mathbf{R}) = \int p(r_{ij}|\boldsymbol{u}_i, \boldsymbol{v}_j) p(\boldsymbol{u}_i, \boldsymbol{v}_j|\mathbf{R}) d\boldsymbol{u}_i d\boldsymbol{v}_j$, which

is in general intractable. Suppose we are given a set of S samples $\{\mathbf{U}^{(s)}, \mathbf{V}^{(s)}\}_{s=1}^{S}$ generated by a Gibbs sampler for this matrix factorization model, where $\mathbf{U}^{(s)} = \{\boldsymbol{u}_i^{(s)}\}_{i=1}^{N}$ and $\mathbf{V}^{(s)} = \{\boldsymbol{v}_j^{(s)}\}_{j=1}^{M}$.

Given these samples, derive the expressions for the sample based approximation of the *mean (expectation)* as well as the *variance* of any entry r_{ij} of the matrix $\mathbf{R}$.

Hint: Note that we can write each r_{ij} as $\boldsymbol{u}_i^\top \boldsymbol{v}_j + \epsilon_{ij}$ where $\epsilon_{ij} \sim \mathcal{N}(\epsilon_{ij}|0, \beta^{-1})$.

Problem 5 (20+20 = 40 marks)

(Part 1: Implementing A Rejection Sampler) Consider a distribution $p(x) \propto \exp(\sin(x))$ for $-\pi \leq x \leq \pi$. Denote $\exp(\sin(x))$ as $\tilde{p}(x)$. Suppose we want to use Rejection Sampling to sample from $p(x)$ and use a proposal distribution $q(x) = \mathcal{N}(x|0, \sigma^2)$. Find out the expression for the optimal value of the constant M such that $Mq(z) \geq \tilde{p}(x)$, as required in Rejection Sampling.

Using this value of M and some suitably chosen σ^2, draw 10,000 samples from $p(x)$ and plot the resulting histogram of the samples. Submit your code in form of a Python notebook.

(Part 2: Implementing MH Sampling for 2-D Gaussian) In this problem, your task is to implement MH sampling to generate random samples from a 2-D Gaussian $p(\boldsymbol{z}) = \mathcal{N}\left(\begin{bmatrix}4\\4\end{bmatrix}, \begin{bmatrix}1 & 0.8\\0.8 & 1\end{bmatrix}\right)$.

To sample from $p(\boldsymbol{z})$, you will use a proposal distribution $q(\boldsymbol{z}^{(t)}|\boldsymbol{z}^{(t-1)}) = \mathcal{N}\left(\boldsymbol{z}^{(t-1)}, \begin{bmatrix}\sigma^2 & 0\\0 & \sigma^2\end{bmatrix}\right)$ and play with different values of the proposal distribution's variance σ^2. For generating the candidate sample from the proposal distribution, you can use existing functions from any Python based library (e.g., numPy), but you must not use the actual distribution $p(\boldsymbol{z})$ to generate the samples.

You will experiment with the following values of $\sigma^2 : 0.01, 1, 100$. For each of these cases, run the MH sampler long enough to collect 10,000 samples and show the plots of the generated samples on a 2-D plane for 100 samples, 1000 samples, and 10,000 samples (similar to the plots of slide 7, lecture-18).

Looking at the plots, which of the 3 proposals ($q(\boldsymbol{z}^{(t)}|\boldsymbol{z}^{(t-1)})$ with $\sigma^2 : 0.01, 1, 100$) seems the best choice to you? What is the rejection rate in each of these cases (rejection rate is the ratio of number of samples rejected and the total number of candidate samples generated)? Submit the code as well as the plots.