Topics in Probabilistic Modeling and Inference (CS698X)
Homework 3 (Due date: April 3, 2018, 11:59pm)

Instructions

- We will only accept electronic submissions and the main writeup must be as a PDF file. If you are handwriting your solutions, please scan the hard-copy and convert it into PDF. Your name and roll number should be clearly written at the top. In case you are submitting multiple files, all files must to be zipped and **submitted as a single file** (named: your-roll-number.zip). Please do not email us your submissions. Your submissions have to be uploaded at the following link: `https://tinyurl.com/y9fw8mu3`).
- Each late submission will receive a 10% penalty per day for up to 3 days. No submissions will be accepted after the 3rd late day.

Problem 1 (20 marks)

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 $v_1 > v_0 > 0$. 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 $\boldsymbol{w}$ (maximum in 50 words)?
- Derive an EM algorithm for doing inference for this model. Your algorithm should give the posterior over the weight vector $\boldsymbol{w}$ and point estimates (MAP) for the remaining unknowns $(\boldsymbol{\gamma}, \sigma^2, \theta)$.

Problem 2 (20 marks)

For the Gaussian mean and precision inference problem, derive the mean-field updates for the variational distribution $q(\mu, \tau)$ using the approach based on writing down the ELBO and explicitly taking the derivatives. Assume we have N observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn i.i.d. from a univariate Gaussian $\mathcal{N}(\mu, \tau^{-1})$ and assume a normal-gamma prior on $p(\mu, \tau)$ as we saw in the class (lecture 16).

Problem 3 (20 marks)

Read the paper "Black Box Variational Inference" by Ranganath et al (2014) and write a summary ($\sim$ 500 words) highlighting the key aspects of this paper. The paper proposes several strategies to reduce the variance of the ELBO's gradients. Your summary must contain a discussion about these strategies. Also discuss any other specific points that you like/dislike about the paper and the methods proposed therein. Feel free to discuss the paper with your classmates but your writeup must be entirely in your own words.

Problem 4 (40 marks)

Consider the model from Problem 6, Homework 2, i.e., $\boldsymbol{x}_n = \mathbf{W}\boldsymbol{z}_n + \epsilon_n$, where $\boldsymbol{x}_n \in \mathbb{R}^D$, $\boldsymbol{z}_n \in \{0,1\}^K$ is a binary vector, and $\mathbf{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K]$ is $D \times K$ matrix. Assume the full generative story to be as follows

$$
\begin{aligned}
\pi_k &\sim \text{Beta}(\alpha/K, 1) \quad \text{for } k = 1, \ldots, K \\
z_{nk} &\sim \text{Bernoulli}(\pi_k) \quad \text{for } n = 1, \ldots, N, k = 1, \ldots, K \\
\boldsymbol{w}_k &\sim \mathcal{N}(0, \sigma_w^2 \mathbf{I}_D) \quad \text{for } k = 1, \ldots, K \\
\boldsymbol{x}_n &\sim \mathcal{N}(\mathbf{W}\boldsymbol{z}_n, \sigma_x^2 \mathbf{I}_D) \quad \text{for } n = 1, \ldots, N
\end{aligned}
$$

Your goal is to derive the mean-field variational inference algorithm to infer the unknowns $\mathbf{Z}, \mathbf{W}, \pi$ of this model. Assume hyperparameters $\alpha, \sigma_w^2, \sigma_x^2$ to be known.

You should use a fully-factorized mean-field assumption, i.e., $q(\mathbf{Z}, \mathbf{W}, \pi) = \prod_{k=1}^{K}[\prod_{n=1}^{N} q(z_{nk})]q(\boldsymbol{w}_k)q(\pi_k)$ and derive each factor of this variational approximation, i.e., $q(z_{nk})$, $q(\boldsymbol{w}_k)$, $q(\pi_k)$, using the "reading off" approach based on the standard form of each mean-field updates, e.g., $\log q(\mathbf{Z}_j) = \mathbb{E}_{i \neq j}[\log p(\mathbf{X}, \mathbf{Z})] + \text{const}$