STA2311 (Fall 2023) - Practice Problems for Class 5 (Variational Inference)

1. Let $p(x,y)$ and $q(x,y)$ be two bivariate mass functions. Write $p_1(x) = \sum_y p(x,y)$ and $p_2^x(y) = p(y \mid x)$, and write $q_1(x)$ and $q_2^x(y)$ similarly. Prove that

$$\mathrm{KL}(p \,||\, q) = \mathrm{KL}(p_1 \,||\, q_1) + \mathbb{E}_X\big[\mathrm{KL}(p_2^X \,||\, q_2^X)\big],$$

   where $X \sim p_1$.

2. Show that the KL-divergence does not always satisfy the triangle inequality; that is, there exist distributions p, q, r such that

$$\mathrm{KL}(p \,||\, r) \not\leq \mathrm{KL}(p \,||\, q) + \mathrm{KL}(q \,||\, r)$$

3. Let $p_1(\boldsymbol{x}) = \mathcal{N}_d(\boldsymbol{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $p_2(\boldsymbol{x}) = \mathcal{N}_d(\boldsymbol{x} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Compute $\mathrm{KL}(p_1 \,||\, p_2)$.

4. Consider the *differential entropy* $H[\cdot]$ defined on the space of density functions.
   (a) Show that the differential entropy is translation invariant in the sense that if $X \sim f$ and $X + c \sim f_c$, then $H[f] = H[f_c]$ for all $c \in \mathbb{R}$.
   (b) Show that among all continuous univariate distributions f with mean μ and variance σ^2, the $\mathcal{N}(\mu, \sigma^2)$ distribution is the one that maximizes $H[f]$.

5. Suppose we approximate a d-dimensional distribution $p(\boldsymbol{z})$ by a mean-field variational family $q(\boldsymbol{z}) = \prod_{i=1}^d q_i(z_i)$. Show that minimizing $\mathrm{KL}(p \,||\, q)$ with respect to one factor $q_i(z_i)$, keeping all other factors fixed, leads to the optimal solution

$$q_i^*(z_i) = \int p(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z}_{-i}.$$

6. Consider linear regression: we have independent observations $Y_1, \ldots, Y_n$ and covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^p$ with $Y_i \mid \boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\beta}^\top \boldsymbol{x}_i, \sigma^2)$ for some $\boldsymbol{\beta} \in \mathbb{R}^p$; we assume that $\sigma^2 > 0$ is known. We adopt a Bayesian model and impose a $\mathcal{N}_p(\boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$ prior on $\boldsymbol{\beta}$ and a $\mathrm{Gamma}(a_0, b_0)$ prior on α. Approximate the posterior $p(\boldsymbol{\beta}, \alpha \mid \boldsymbol{y})$ by deriving a mean-field variational approximation of the form $q(\boldsymbol{\beta}, \alpha) = q(\boldsymbol{\beta}) \cdot q(\alpha)$.

7. Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathrm{Gamma}(\theta, 1)$ for some $\theta > 0$, and let $Z_i \mid X_i = x_i \overset{\text{indep}}{\sim} \mathrm{Gamma}(x_i, \theta)$. We are once again Bayesians and adopt a $\mathrm{Gamma}(a, b)$ prior on θ. Approximate the posterior $p(\theta, \boldsymbol{z} \mid \boldsymbol{x})$ by deriving a mean-field variational approximation of the form $q(\theta, \boldsymbol{z}) = q(\theta) \cdot q(\boldsymbol{z})$.