Advanced Machine Learning – Deep Generative Models Exercise Sheet 02

Variational Inference

Problem 1: Consider the following latent variable model.

$$p_\theta(z) = \mathrm{Exp}(z|\theta) = \begin{cases} \theta \exp(-\theta z) & \text{if } z > 0, \\ 0 & \text{else.} \end{cases}$$

$$p(x|z) = \mathcal{N}(x|z, 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-z)^2\right),$$

where $x \in \mathbb{R}$ is the observed data and $z \in \mathbb{R}_+$ is the latent variable. We have observed a single data point x and now would like to maximize the marginal log-likelihood $\log p_\theta(x) = \log\left(\int p(x|z)p_\theta(z)dz\right)$ w.r.t. the model parameters $\theta \in \mathbb{R}_+$. For this we will use variational inference.

We define the following parametric family of variational distributions

$$q_\phi(z) = \mathrm{Exp}(z|\phi) = \begin{cases} \phi \exp(-\phi z) & \text{if } z > 0 \\ 0 & \text{else;} \end{cases}$$

that is parametrized by $\phi \in \mathbb{R}_+$. We are interested in solving the optimization problem

$$\max_{\theta>0, \phi>0} \mathcal{L}(\theta, \phi).$$

a) Assume that θ is known and fixed. Does there exist a value of ϕ such that the ELBO is tight, i.e. $\log p_\theta(x) = \mathcal{L}(\theta, \phi)$? Justify your answer.

We remember from the lecture that

$$\log p_\theta(x) = \mathcal{L}(\theta, \phi) + \mathbb{KL}(q_\phi(z) \| p_\theta(z|x))$$

For a fixed θ, if we find some value of ϕ such that $\mathbb{KL}(q_\phi(z) \| p_\theta(z|x)) = 0$, then we'll have

$$\log p_\theta(x) = \mathcal{L}(\theta, \phi)$$

We also remember from the lecture that $\mathbb{KL}(q_\phi(z) \| p_\theta(z|x)) = 0$ only holds if $q_\phi(z) \equiv p_\theta(z|x)$. Therefore, the original question can be reformulated as

"Does there exist a value ϕ, such that $q_\phi(z) = p_\theta(z|x)$ for all z"?

To answer this question, we look at the unnormalized posterior over z

$$\begin{aligned} p_\theta(z|x) &\propto p(x|z)p_\theta(z) \\ &\propto \exp\left(-\frac{1}{2}(x-z)^2\right) \exp(-\theta z)\mathbf{1}(z > 0) \\ &= \exp\left(-\frac{1}{2}x^2 + xz - \frac{1}{2}z^2 - \theta z\right)\mathbf{1}(z > 0) \\ &\propto \exp\left(-\frac{1}{2}z^2 + (x-\theta)z\right)\mathbf{1}(z > 0) \end{aligned}$$

Here, $\mathbf{1}(\cdot)$ is the indicator function.

$$\mathbf{1}(\text{condition}) = \begin{cases} 1 & \text{if condition is true,} \\ 0 & \text{else.} \end{cases}$$

We absorb the terms that don't depend on z into the $\propto$ sign since we only care about the distribution over z.

Now, let's have a look at our approximate posterior $q_\phi(z)$

$$q_\phi(z) \propto \exp(-\phi z)\mathbf{1}(z > 0)$$

No matter which value of ϕ we choose, it cannot happen that

$$-\phi z = -\frac{1}{2}z^2 + (x - \theta)z$$

because we have a term quadratic in z on the right hand side. Hence, we conclude that for any $\phi \in \mathbb{R}_+$ it holds that $\mathbb{KL}(q_\phi(z) \| p_\theta(z|x)) > 0$, and therefore $\log p_\theta(x) > \mathcal{L}(\theta, \phi)$.

b) Write down the ELBO $\mathcal{L}(\theta, \phi)$ for the above probabilistic model $p_\theta(x, z)$ and the variational distribution $q_\phi(z)$ and simplify it as far as you can. Your final answer should be a closed-form expression (no integrals or expectations).

By definition, the ELBO is equal to

$$\begin{aligned} \mathcal{L}(\theta, \phi) &= \mathop{\mathbb{E}}_{z \sim q_\phi(z)} [\log p(x|z) + \log p_\theta(z) - \log q_\phi(z)] \\ &= \mathop{\mathbb{E}}_{z \sim q_\phi(z)} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}(x - z)^2 + \log\theta - \theta z - \log\phi + \phi z \right] \\ &= \mathop{\mathbb{E}}_{z \sim q_\phi(z)} \left[-\frac{1}{2}z^2 + xz + \log\theta - \theta z - \log\phi + \phi z \right] + \text{const.} \end{aligned}$$

From the properties of the exponential distribution (`https://en.wikipedia.org/wiki/Exponential_distribution`) we know that $\mathbb{E}_{z \sim q_\phi(z)}[z] = \frac{1}{\phi}$ and $\mathbb{E}_{z \sim q_\phi(z)}[z^2] = \frac{2}{\phi^2}$

$$\begin{aligned} &= -\frac{1}{\phi^2} + \frac{x}{\phi} + \log\theta - \frac{\theta}{\phi} - \log\phi + 1 + \text{const.} \\ &= -\frac{1}{\phi^2} + \frac{x - \theta}{\phi} + \log\theta - \log\phi + \text{const.} \end{aligned}$$

Note that our distribution $q_\phi(z)$ can only produce positive values of z, so we don't have to worry about what happens when $z \leq 0$.

c) Compute the gradients of the ELBO $\nabla_\theta \mathcal{L}(\theta, \phi)$ and $\nabla_\phi \mathcal{L}(\theta, \phi)$.

We simply need to compute the derivatives of the expression obtain in (b) w.r.t. θ and ϕ and obtain

$$\frac{\partial}{\partial\theta}\mathcal{L}(\theta,\phi) = -\frac{1}{\phi} + \frac{1}{\theta}$$
$$\frac{\partial}{\partial\phi}\mathcal{L}(\theta,\phi) = \frac{2}{\phi^3} - \frac{x-\theta}{\phi^2} - \frac{1}{\phi}$$

Problem 2: You want to draw samples from an exponential distribution with rate ϕ with reparametrization. Assume that

$$q_\phi(z) = \mathrm{Exp}(z|\phi) = \begin{cases} \phi\exp(-\phi z) & \text{if } z > 0 \\ 0 & \text{else;} \end{cases}$$

where $\phi \in \mathbb{R}_+$.

a) You have access to an algorithm that produces samples ϵ from an exponential distribution with unit rate, that is

$$b(\epsilon) = \mathrm{Exp}(\epsilon|1) = \begin{cases} \exp(-\epsilon) & \text{if } \epsilon > 0 \\ 0 & \text{else.} \end{cases}$$

Find a deterministic transformation $T_\phi\colon \mathbb{R}_+ \to \mathbb{R}_+$ that converts a sample $\epsilon \sim b(\epsilon)$ into a sample from $q_\phi(z)$.

We can find the transformation T_ϕ that transforms $\epsilon \sim \mathrm{Exp}(1)$ into $z = T_\phi(\epsilon) \sim \mathrm{Exp}(\phi)$ by considering the cumulative distribution functions (CDFs) of the two random variables. We denote the CDF of ϵ as $F_\epsilon(a) = \Pr(\epsilon \le a)$, and the CDF of z as $F_z(a) = \Pr(z \le a)$.

Our goal is to find a transformation T_ϕ such that

$$\begin{aligned} F_z(a) &= \Pr(z \le a) \\ &= \Pr(T_\phi(\epsilon) \le a) \\ &= \Pr(\epsilon \le T_\phi^{-1}(a)) \\ &= F_\epsilon(T_\phi^{-1}(a)) \end{aligned}$$

From the properties of the exponential distribution we know that

$$F_z(a) = 1 - \exp(-\phi a) \qquad \text{and} \qquad F_\epsilon(T_\phi^{-1}(a)) = 1 - \exp(-T_\phi^{-1}(a))$$

This means that

$$T_\phi^{-1}(a) = \phi a \qquad \Longrightarrow \qquad T_\phi(x) = \frac{1}{\phi}x$$

That is, $z = T_\phi(\epsilon) = \frac{1}{\phi}\epsilon$ is the desired transformation.

b) Now, you have access to an algorithm that produces samples u from a uniform distribution on $[0, 1]$, that is

$$b(u) = \begin{cases} 1 & \text{if } u \in [0,1] \\ 0 & \text{else.} \end{cases}$$

Find a deterministic transformation $S_\phi \colon [0,1] \to \mathbb{R}_+$ that converts a sample $u \sim b(u)$ into a sample $z = S_\phi(u) \sim q_\phi(z)$.

We can reuse the result from part (a) here and simply look for a transformation $R \colon [0,1] \to \mathbb{R}_+$ that converts $u \sim U([0,1])$ into a sample from $\epsilon \sim \text{Exp}(1)$. Then we can convert $\epsilon = R(u) \sim \text{Exp}(1)$ into $z \sim \text{Exp}(\phi)$ using the transformation T_ϕ that we found in part (a) as $z = S_\phi(u) = T_\phi(R(u))$.

We again use the CDFs $F_u(a) = \Pr(u \leq a)$ and $F_\epsilon(a) = \Pr(\epsilon \leq a)$. Using similar reasoning as before, we need to satisfy

$$\begin{aligned} 1 - \exp(-a) &= F_\epsilon(a) \\ &= \Pr(\epsilon \leq a) \\ &= \Pr(R(u) \leq a) \\ &= \Pr(u \leq R^{-1}(a)) \\ &= F_u(R^{-1}(a)) \\ &= R^{-1}(a) \end{aligned}$$

We see that $R(x) = -\log(1-x)$ satisfies the above equation.

Therefore, $S_\phi(x) = T_\phi(R(u)) = -\frac{\log(1-u)}{\phi}$ is the desired transformation.

Problem 3: You are given two distributions $q(\boldsymbol{z})$ and $p(\boldsymbol{z})$ over some random vector $\boldsymbol{z} \in \mathbb{R}^D$. Assume that both distributions can be factorized as

$$q(\boldsymbol{z}) = \prod_{i=1}^{D} q_i(z_i) \qquad\qquad p(\boldsymbol{z}) = \prod_{i=1}^{D} p_i(z_i).$$

(This is equivalent to saying that each component z_i is independent of z_j for $j \neq i$ under the distributions q and p). Your task is to prove that in this case the following equality holds

$$\mathbb{KL}(q(\boldsymbol{z}) \| p(\boldsymbol{z})) = \sum_{i=1}^{D} \mathbb{KL}(q_i(z_i) \| p_i(z_i)).$$

$$\mathbb{KL}(q(\boldsymbol{z})\|p(\boldsymbol{z})) = \int q(\boldsymbol{z}) \log \frac{q(\boldsymbol{z})}{p(\boldsymbol{z})} d\boldsymbol{z} \tag{1}$$

$$= \int \cdots \int q(z_1, ..., z_D) \log \frac{q(z_1, ..., z_D)}{p(z_1, ..., z_D)} dz_1 ... dz_D \tag{2}$$

$$= \int \cdots \int q_1(z_1) \cdots q_D(z_D) \log \left(\prod_{i=1}^{D} \frac{q_i(z_i)}{p_i(z_i)} \right) dz_1 ... dz_D \tag{3}$$

$$= \sum_{i=1}^{D} \left(\int \cdots \int q_1(z_1) \cdots q_D(z_D) \log \frac{q_i(z_i)}{p_i(z_i)} dz_1 ... dz_D \right) \tag{4}$$

$$= \sum_{i=1}^{D} \left(\int q_i(z_i) \log \frac{q_i(z_i)}{p_i(z_i)} \underbrace{\left(\int \cdots \int q_1(z_1) \cdots q_{i-1}(z_{i-1}) q_{i+1}(z_{i+1}) \cdots q_D(z_D) dz_1 ... dz_{i-1} dz_{i+1} ... dz_D \right)}_{=1} dz_i \right) \tag{5}$$

$$= \sum_{i=1}^{D} \left(\int q_i(z_i) \log \frac{q_i(z_i)}{p_i(z_i)} dz_i \right) \tag{6}$$

$$= \sum_{i=1}^{D} \mathbb{KL}(q_i(z_i)\|p_i(z_i)) \tag{7}$$

Here, we used the following properties:

- Lines 1-2: $q(\boldsymbol{z}) = q(z_1, ..., z_D)$ is just another way of writing the same thing.
- Lines 2-3: Distributions $q(\boldsymbol{z})$ and $p(\boldsymbol{z})$ factorize (from the problem statement).
- Lines 3-4: $\log \left(\prod_i x_i \right) = \sum_i \log x_i$ and "integral of a sum = sum of integrals".
- Lines 4-5: We can change the order in which we compute the integrals.
- Lines 5-6: $\int q_j(z_j) dz_j = 1$ for every j since each q_j is a valid probability density.
- Lines 6-7: Use the definition of KL divergence.