Probabilistic Programming Homework 4

Submit your solutions to the TA in the homework submission box in the third floor of the E3-1 building by 2:00pm on 1 June 2018 (Friday). If you type up your solutions, you can email them to him (kwonsoo.chae@gmail.com).

Question 1

This question is about amortised inference. Consider finite sets X and Y that are ranged over by x and y, respectively. Assume that we are given a probability distribution $p(x, y)$ on $X \times Y$ in the form of $p(x)$ and $p(y|x)$. The x part of (x, y) is an unobserved latent state, and the y part is an observation. We consider a situation where we want to perform posterior inference on $p(x, y)$ multiple times for different observations $y_1, y_2, \ldots, y_m$. The idea of amortised inference is to do certain preprocessing, which takes some time now but will save time in the future when we carry out these repeated inference tasks with observations $y_1, \ldots, y_m$.

More concretely, in the amortised inference, we consider a conditional proposal distribution $q_\theta(x|y)$ parameterised by $\theta \in \mathbb{R}^k$ such that for fixed x and y, the function $\theta \mapsto q_\theta(x|y)$ is a differentiable function from $\mathbb{R}^k$ to the interval $[0, 1]$. We often construct such a proposal using a neural net, in which case θ is the weights of the neural net. Given such a parameterised conditional proposal, the amortised inference works as follows. First, it solves the following optimisation problem:

$$\operatorname{argmin}_\theta \mathbb{E}_{p(y)} \left[\mathrm{KL}\Big(p(x|y) \Big|\Big| q_\theta(x|y) \Big) \right] \tag{1}$$

where

$$\mathrm{KL}\Big[p(x|y) \,\Big|\Big|\, q_\theta(x|y) \Big] = \sum_x \left(p(x|y) \cdot \log \frac{p(x|y)}{q_\theta(x|y)} \right).$$

Solving this optimisation problem corresponds to the preprocessing mentioned before. Let θ^* be the solution of this optimisation problem. Second, when we are given an observation y_i and asked to estimate the posterior $p(x|y_i)$, we instantiate $q_{\theta^*}(x|y_i)$ by y_i, and perform the importance sampling using $q_{\theta^*}(x|y_i)$ as a proposal.

Of course, the most challenging part of this amortised inference is to solve the optimisation problem (1). A common approach is to use gradient descent (which by the way interacts very well with the backpropagation algorithm for neural nets).

(a) Prove that the optimisation problem in (1) is equivalent to the following problem:

$$\operatorname{argmin}_\theta \left[\mathrm{KL}\Big(p(x, y) \Big|\Big| p(y) \cdot q_\theta(x|y) \Big) \right].$$

(b) Prove the following key equation that lies behind the gradient-based algorithm:

$$\nabla_\theta \left(\mathrm{KL}\Big[p(x, y) \,\Big|\Big|\, p(y) \cdot q_\theta(x|y) \Big] \right) = \mathbb{E}_{p(x,y)} \left[-\nabla_\theta \log(q_\theta(x|y)) \right]. \tag{2}$$

From this equation, derive a gradient-descent algorithm that approximately solves the problem (1). Your algorithm does not have to be super efficient.

(c) Suppose now that we are given a distribution $p_D(y)$ on observed data. Typically, $p_D(y)$ is defined to be a uniform distribution over a finite collection (strictly speaking multiset) of observations $\{y_1, \ldots, y_m\}$. This time we want to do amortised inference using $p_D(y)$ instead of the marginal distribution $p(y)$ from the model. That is, we want to solve the following optimisation problem:

$$\operatorname{argmin}_\theta \left[\mathrm{KL}\Big(p_D(y) \cdot p(x|y)\Big|\Big|p_D(y) \cdot q_\theta(x|y)\Big)\right], \tag{3}$$

which is equivalen to

$$\operatorname{argmin}_\theta \mathbb{E}_{p_D(y)} \left[\mathrm{KL}\Big(p(x|y)\Big|\Big|q_\theta(x|y)\Big)\right].$$

As before, we solve this optimisation problem using gradient descent. We can use the following estimator for the gradient:

$$\nabla_\theta \,\mathrm{KL}\Big(p_D(y) \cdot p(x|y)\Big|\Big|p_D(y) \cdot q_\theta(x|y)\Big) \;\approx\; -\frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sum_{j=1}^{M} w_{i,j}} \cdot \sum_{j=1}^{M} w_{i,j} \cdot \nabla_\theta \,\log q_\theta(x_{i,j}|y_i).$$

Here (i) $y_1, \ldots, y_N$ are independent samples from p_D, (ii) $x_{i,1}, \ldots, x_{i,M}$ are independent samples from $q_\theta(x_{i,j}|y_i)$ for each y_i, and (iii) $w_{i,j}$ is defined by

$$w_{i,j} = \frac{p(x_{i,j}, y_i)}{q_\theta(x_{i,j}|y_i)}.$$

This estimator is a part of Bornschein and Bengio's reweighted wake-sleep algorithm in their ICLR'15 paper.

Explain why the above estimator gives a reasonable approximation of the gradient. Here I am not asking for a proof, but a semi-rigorous justification (which can be turned into a proof of so called consistency of the estimator). Hint: My answer for this question uses the justification of (self-normalising) importance sampling that we discussed during lecture.