Evidence lower bounds with built-in baselines

Justin

December 23, 2022

1 Problem setup

The goal is to maximize the log marginal likelihood,

$$\log p(x) = \log \sum_z p(x, z), \tag{1}$$

for a latent variable model. The derivative of equation 1 is

$$\nabla \log \sum_z p(x, z) = \frac{p(x, z)}{p(x)} \nabla \log p(x, z) = p(z \mid x) \nabla \log p(x, z), \tag{2}$$

the expected gradient under the posterior. When exact marginalization and exact computation of the posterior is intractable,[1] a common approach is to introduce a variational approximation the the posterior $q(z \mid x)$ and optimize the following lower bound

$$\log p(x) = \log \sum_z q(z \mid x) \frac{p(x, z)}{q(z \mid x)} \geq \sum_z q(z \mid x) \log \frac{p(x, z)}{q(z \mid x)}.^2 \tag{3}$$

Empirically, we find that directly optimizing equation 3 to be more difficult than optimizing the marginal likelihood (equation 1), often requiring a baseline:

$$\begin{aligned} &\nabla_q \sum_z q(z \mid x) \log \frac{p(x, z)}{q(z \mid x)} \\ &= \nabla_q \sum_z q(z \mid x) \log p(x \mid z) - KL[q(z \mid x) || p(z)] \\ &= \nabla_q \sum_z q(z \mid x)(\log p(x \mid z) - B) - KL[q(z \mid x) || p(z)], \end{aligned} \tag{4}$$

where the baseline B is not a function of z.[3] Computing this baseline B can be relatively expensive, requiring multiple evaluations of $p(x \mid z)$. A common choice of baseline is the sample-average baseline, where $B = \sum_{z' \in Z} \log p(x \mid z')$ and Z is a set of iid samples.[4]

We show that the gradient of the marginal likelihood (equation 2) already contains a built-in baseline, and use that to derive a simple and computable lower bound of the marginal likelihood (equation 1) whose gradient does not require the manual engineering of a baseline.

[1] Marginalization and computation of the posterior take the same amount of computation.

[2] The gap between the two is given by $KL[q(z \mid x) || p(z \mid x)]$.

[3] The derivative is a linear operator, and $\nabla \sum_z q(z \mid x) B = B \nabla \sum_z q(z \mid x) = B \cdot 0$.

[4] A more efficient alternative is the leave-one-out baseline, which is efficient if the results of $p(x \mid z)$ are already available for a set of iid z.

2 The gradient of logsumexp gives the scaled regret

The key component of the marginal likelihood is the logsumexp operation, a smooth version of max: $\log \sum_z \exp f(z)$. Similar to the gradient of the marginal likelihood, the gradient of logsumexp is given by

$$\nabla \log \sum_{z' \in Z} \exp f(z') = \frac{e^{f(z)}}{\sum_{z' \in Z} e^{f(z')}} \nabla f(z) = \underbrace{e^{f(z) - \log \sum_{z' \in Z} e^{f(z')}}}_{R} \nabla f(z) \approx e^{f(z) - \max_{z' \in Z} f(z')} \nabla f(z).$$

The last approximation holds because logsumexp is a smooth approximation of max. We can therefore think of R as a built-in baseline. R is also the exponentiated regret, e.g. difference between a given $f(z)$ and the best $f(z)$. The gradient has another nice property: it is always scaled between 0 and 1.