Bayes by backprop

Apostolos Psaros

1 Introduction

Suppose we have only one weight w and the posterior is approximated by $q_\theta(w)$, parametrized by θ (also one-dimensional). The ELBO is given as

$$ELBO(\theta) = \int \log\left(p(\mathcal{D}|w,\mathcal{H})\right) q_\theta(w)dw - KL\left(q_\theta(w)||p(w|\mathcal{H})\right) \tag{1}$$

For maximization of the ELBO we need to differentiate with respect to θ. In these notes, we will focus on the derivative of the integral term, i.e.,

$$I(\theta) = \frac{\partial}{\partial\theta}\int f(w)q_\theta(w)dw = \int f(w)\frac{\partial}{\partial\theta}q_\theta(w)dw \tag{2}$$

where, for notation convenience, $\log\left(p(\mathcal{D}|w,\mathcal{H})\right)$ is replaced by $f(w)$ and the fact that $f(w)$ is independent of θ has been used. How can we approximate $I(\theta)$?

2 Monte Carlo estimators

In general, an integral can be approximated by a Monte Carlo (MC) estimator as

$$\int f(w)p(w)dw \approx \frac{\sum_{j=1}^{M} f(\hat{w}_j)}{M} \tag{3}$$

where $\{\hat{w}_j\}_{j=1}^{M}$ are M samples from the distribution $p(w)$. For $M=1$

$$\int f(w)p(w)dw \approx f(\hat{w}),\ \hat{w}\sim p(w) \tag{4}$$

This estimator is unbiased if the M samples are independent and the estimator variance depends on M and on the form of f.

3 Integral approximation

Suppose that we approximate the first integral in Eq. (2) as

$$\int f(w)q_\theta(w)dw \approx f(\hat{w}),\ \hat{w}\sim q_\theta(w) \tag{5}$$

Then we cannot take the derivative with respect to θ. On the other hand, if we use the second integral of Eq. (2), then we cannot use the MC estimator of Eq. (4).

One workaround is to express Eq. (2) equivalently as

$$I(\theta) = \int f(w) \frac{\partial \log q_\theta(w)}{\partial \theta} q_\theta(w) dw \tag{6}$$

by using $\frac{\partial}{\partial \theta} q_\theta(w) = \frac{\partial \log q_\theta(w)}{\partial \theta} q_\theta(w)$. Then we can approximate $I(\theta)$ by

$$\hat{I}(\theta) = f(\hat{w}) \frac{\partial \log q_\theta(\hat{w})}{\partial \theta},\ \hat{w} \sim q_\theta(w) \tag{7}$$

This is called score function (or likelihood ratio) estimator.

Another estimator arises by using the reparametrization trick. In general, we are trying to derive an MC estimator that involves drawing samples of w from $q_\theta(w)$. At the same time, we are trying to differentiate with respect to θ. In this regard, the reparametrization trick separates the randomness involved in the distribution $q_\theta(w)$ by the functional form of $q_\theta(w)$, which relates to θ. Differentiating and drawing samples become disentangled. Specifically, for some transformation $w = g(\theta, \epsilon)$, where ϵ is the random part, we express $q_\theta(w)$ as

$$q_\theta(w) = \int q_\theta(w, \epsilon) d\epsilon = \int q_\theta(w|\epsilon) p(\epsilon) d\epsilon \tag{8}$$

where $q_\theta(w|\epsilon) = \delta(w - g(\theta, \epsilon))$. Therefore,

$$\int f(w) q_\theta(w) dw = \int \int f(w) \delta(w - g(\theta, \epsilon)) p(\epsilon) d\epsilon dw = \int f(g(\theta, \epsilon)) p(\epsilon) d\epsilon \tag{9}$$

and

$$I(\theta) = \frac{\partial}{\partial \theta} \int f(w) q_\theta(w) dw = \int \frac{d}{dw} f(g(\theta, \epsilon)) \frac{\partial}{\partial \theta} g(\theta, \epsilon) p(\epsilon) d\epsilon \tag{10}$$

Then we can approximate $I(\theta)$ by

$$\hat{I}(\theta) = \frac{\partial}{\partial \theta} f(g(\theta, \hat{\epsilon})) = f'(g(\theta, \hat{\epsilon})) \frac{\partial}{\partial \theta} g(\theta, \hat{\epsilon}),\ \hat{\epsilon} \sim p(\epsilon) \tag{11}$$

This is known as path-wise estimator.

4 Comparison with MAP

In Bayes by backprop we optimize for θ and the update is given by

$$\Delta\theta = \frac{\partial}{\partial \theta} \log \left(p(\mathcal{D}|g(\theta, \hat{\boldsymbol{\epsilon}}), \mathcal{H})\right) - \frac{\partial}{\partial \theta} KL \left(q_\theta(w) || p(w|\mathcal{H})\right) \tag{12}$$

where $\hat{\boldsymbol{\epsilon}}$ is a vector of N random samples from $p(\epsilon)$; one for each datapoint in $\mathcal{D}$. In MAP gradient descent we optimize for w and the update is given by

$$\Delta w = \frac{\partial}{\partial w} \log \left(p(\mathcal{D}|w, \mathcal{H})\right) + \frac{\partial}{\partial w} \log \left(p(w|\mathcal{H})\right) \tag{13}$$

Note that the MAP update can be obtained as a special case of Bayes by backprop. Specifically, if we set $q_\theta(w) = \delta(w - \theta)$, then $g(\theta, \epsilon) = \theta$ and

$$\begin{aligned} KL\left(q_\theta(w)||p(w|\mathcal{H})\right) &= \int \log\left(\frac{q_\theta(w)}{p(w|\mathcal{H})}\right) q_\theta(w) dw \\ &= \int \delta(w-\theta)\log\delta(w-\theta)dw - \int \delta(w-\theta)\log p(w|\mathcal{H})dw \\ &= -\log p(\theta|\mathcal{H}) \end{aligned} \tag{14}$$

5 Multi-dimensional case

The update for multi-dimensional $\boldsymbol{w}$ and $\boldsymbol{\theta}$ and Bayes by backprop SGD is given by

$$\Delta\boldsymbol{\theta} = \nabla_{\boldsymbol{\theta}} \sum_{i\in S} \frac{N}{|S|} \log p(y_i|\boldsymbol{g}(\boldsymbol{\theta}, \hat{\boldsymbol{\epsilon}}_i), x_i, \mathcal{H}) - \nabla_{\boldsymbol{\theta}} KL\left(q_{\boldsymbol{\theta}}(\boldsymbol{w})||p(\boldsymbol{w}|\mathcal{H})\right) \tag{15}$$

where $\hat{\boldsymbol{\epsilon}}_i$ is a sampled vector from $p(\boldsymbol{\epsilon})$. The update for MAP SGD is given by

$$\Delta\boldsymbol{w} = \nabla_{\boldsymbol{w}} \sum_{i\in S} \frac{N}{|S|} \log p(y_i|\boldsymbol{w}, x_i, \mathcal{H}) + \nabla_{\boldsymbol{w}} \log\left(p(\boldsymbol{w}|\mathcal{H})\right) \tag{16}$$

6 Factorized Gaussian posterior

A common choice for $q_{\boldsymbol{\theta}}(\boldsymbol{w})$ is a factorized Gaussian posterior given as

$$q_{\boldsymbol{\theta}}(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, diag(\boldsymbol{\sigma}^2)) = \prod_{j=1}^{k} \mathcal{N}(w_j|\mu_j, \sigma_j^2) \tag{17}$$

An approximation in which the variational distribution factorizes over the parameters is called mean-field approximation. The reparametrization for this case is given as

$$\boldsymbol{w} = \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{\epsilon}) = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon} \tag{18}$$

and $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\sigma}\}$.