Understanding Reparametrisation Gradients

Philip Schulz and Wilker Aziz

last modified: September 25, 2017

Abstract

This note explains derives in detail the reparametrisation trick presented in Kingma and Welling (2013); Rezende et al. (2014); Titsias and Lázaro-Gredilla (2014). Our derivation mostly follows Titsias and Lázaro-Gredilla (2014). It also gives some advice how terms such as Jacobians should be distributed.

1 Derivation

We assume a model of some data y whose log-likelihood is given as

$$\log p(y|\theta) = \int \log \underbrace{p(y, x|\theta)}_{g(x)} \mathrm{dx} \tag{1}$$

where x is any set of latent variables and θ are the model parameters. The joint likelihood, which we abbreviate as $g(x)$, can be arbitrarily complex; in particular, it can be given by a neural network. Since for comples models exact integration over the latent space is impossible, we employ variational inference for parameter optimisation. The variational parameters are called λ. The objective is

$$\arg\max_{\lambda} = \mathbb{E}_{q(x|\lambda)}\left[\log \frac{g(x)}{q(x|\lambda)}\right] \ . \tag{2}$$

We further assume that exact integration is not possible even under the variational approximation (this is the case in non-conjugate models, such as neural networks). Instead we want to sample gradient estimates using Monte Carlo (MC) sampling. Unfortunately, the MC estimator is not differentiable. We thus to sample the simpler random variable $Z \sim \mathcal{N}(0, I)$ which does not depend on the variational parameters. In order to express $q(\theta|\lambda)$ in terms of the density $\phi(z)$ we need to tranform one into the other.

$$Z = h(x, \lambda) = (X - \mu)C^{-1} \tag{3}$$

$$X = h^{-1}(z, \lambda) = ZC + \mu \tag{4}$$

We thus get the following transformed density which is equivalent to $q(\theta|\lambda)$, where we use $|\cdot|$ to denote the absolute value function.

$$\phi((x-\mu)C^{-1})|\det\frac{dh(x,\lambda)}{x}| = \phi((x-\mu)C^{-1})|C^{-1}| = q(\theta|\lambda) \tag{5}$$

Using the transformation to rewrite the expectation from Equation (2).

$$\mathbb{E}_{q(x|\lambda)}\left[\log\frac{g(x)}{q(x|\lambda)}\right] = \int q(x|\lambda)\log\frac{g(x)}{q(x|\lambda)}\mathrm{dx} \tag{6}$$

$$= \int \phi((X-\mu)C^{-1})|C^{-1}|\log\frac{g(x)}{\phi((x-\mu)C^{-1})|C^{-1}|}\mathrm{dzC} \tag{7}$$

$$= \int \phi((X-\mu)C^{-1})\log\frac{g(x)}{\phi((x-\mu)C^{-1})|C^{-1}|}\mathrm{dz} \tag{8}$$

$$= \int \phi((X-\mu)C^{-1})\log\frac{g((x-\mu)C^{-1})}{\phi((x-\mu)C^{-1})|C^{-1}|}\mathrm{dz} \tag{9}$$

$$= \mathbb{E}_{\phi(z)}\left[\log\frac{\overbrace{g((x-\mu)C^{-1})}^{g(x)}}{\underbrace{\phi((x-\mu)C^{-1})|C^{-1}|}_{q(\theta|\lambda)}}\right] \tag{10}$$

2 Noteworthy Points

- The cancellation of the absolute value of the Jacobian determinant and
- We can usually rewrite $g(x) = p(y|x,\theta)p(x|\theta)$. This enables us to split up the objective function as

$$\mathbb{E}_{q(x|\lambda)}\left[\log\frac{p(y|x,\theta)p(x|\theta)}{q(x|\lambda)}\right] = \mathbb{E}_{q(x|\lambda)}\left[\log p(y|x,\theta)\right] - \mathrm{KL}\left(q(x|\lambda)\;||\;p(x|\theta)\right) \tag{11}$$

In case we can compute the KL term analytically, we do not need to included the Jacobian in the objective.

References

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. 2013. URL http://arxiv.org/abs/1312.6114.

Danilo J. Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1278–1286, 2014. URL http://jmlr.org/proceedings/papers/v32/rezende14.pdf.

Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1971–1979, 2014. URL http://jmlr.org/proceedings/papers/v32/titsias14.pdf.