Understanding Reparametrisation Gradients

Philip Schulz and Wilker Aziz

last modified: October 27, 2017

Abstract

This note explains in detail the reparametrisation trick presented in **???**. Our derivation mostly follows **?**. 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

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

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 assume that the random variable X can be represented by transforming samples from a standard distribution $\phi(z)$ using an affine transformation:

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

$$x = h^{-1}(z,\lambda) = \mu + Cz \ . \tag{3b}$$

Note that $\phi(z)$ does not depend on λ which were absorbed in the affine transformation—this in fact restricts the class of approximations $q(x|\lambda)$ to location-scale distributions.

For the sake of generality we take x and z to be vector valued. Then we write $J_{h(x,\lambda)}$ to denote the Jacobian matrix of the transformation $h(x,\lambda)$, and $J_{h^{-1}(z,\lambda)}$ to denote the Jacobian matrix of the inverse transformation.[1] An important property, which we will use to derive reparameterised gradients, is that the inverse of a Jacobian matrix is related to the Jacobian matrix of the inverse function by $J_{h^{-1}} \circ h(x) = J^{-1}_{h(x)}$.[2]

For an invertible transformation of random variables, it holds that

$$q(x|\lambda) = \phi(h(x,\lambda))\left|\det J_{h(x,\lambda)}\right| \tag{4}$$

[1] Recall that a Jacobian matrix $\mathbf{J} \triangleq J_{f(x)}$ of some vector value function $f(x)$ is such that $J_{i,j} = \frac{\partial}{\partial x_j} f_i(x)$.

[2] The notation $J_{h^{-1}} \circ h(x)$ denotes function composition, that is, $J_{h^{-1}}(z = h(x))$ or equivalently $J_{h^{-1}}(z)\big|_{z=h(x)}$.

and therefore for the transformation in (3) we can write

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

and

$$\phi(z) = q(\mu + Cz|\lambda)|\det C| \ . \tag{6}$$

Re-writing the expectation from Equation (2) in terms of the transformed random variable we have

$$\int q(x|\lambda)\log\frac{g(x)}{q(x|\lambda)}\mathrm{d}x \tag{7a}$$

$$= \int \phi(\underbrace{h(x,\lambda)}_{z})\big|\det J_{h(x,\lambda)}\big| \log\frac{g(x)}{\phi(h(x,\lambda))\big|\det J_{h(x,\lambda)}\big|}\mathrm{d}x \tag{7b}$$

$$= \int \phi(z)\big|\det J_h \circ h^{-1}(z,\lambda)\big| \log\frac{g(h^{-1}(z,\lambda))}{\phi(z)|\det J_h \circ h^{-1}(z,\lambda)|}\big|\det J_{h^{-1}(z,\lambda)}\big|\mathrm{d}z \tag{7c}$$

$$= \int \phi(z)\Big|\det J^{-1}_{h^{-1}(z,\lambda)}\Big| \log\frac{g(h^{-1}(z,\lambda))}{\phi(z)\Big|\det J^{-1}_{h^{-1}(z,\lambda)}\Big|}\big|\det J_{h^{-1}(z,\lambda)}\big|\mathrm{d}z \tag{7d}$$

$$= \int \phi(z)\frac{1}{|\det J_{h^{-1}}(z,\lambda)|} \log\frac{g(h^{-1}(z,\lambda))|\det J_{h^{-1}}(z,\lambda)|}{\phi(z)}|\det J_{h^{-1}}(z,\lambda)|\mathrm{d}z \tag{7e}$$

$$= \int \phi(z)\log\frac{g(h^{-1}(z,\lambda))|\det J_{h^{-1}}(z,\lambda)|}{\phi(z)}\mathrm{d}z \tag{7f}$$

$$= \int \phi(z)\log\left(g(h^{-1}(z,\lambda))\left|\det \underbrace{J_{h^{-1}}(z,\lambda)}_{C}\right|\right)\mathrm{d}z - \int \phi(z)\log\phi(z)\mathrm{d}z \tag{7g}$$

$$= \mathbb{E}_{\phi(Z)}[\log g(h^{-1}(Z,\lambda))] + \log|C| + \mathbb{H}[\phi(Z)] \tag{7h}$$

for which we can easily construct gradient estimates by MC sampling.

A digest of what happened

- In (7b) we applied a change of density.
- In (7c) we applied a change of variable thus expressing every integrand as a function of z rather than x. First, note that this calls for a change of infinitesimal volumes, i.e. $\mathrm{d}x = |\det J_{h^{-1}}(z,\lambda)|\mathrm{d}z$. Second, note that, to express the Jacobian $J_{h(x,\lambda)}$ as a function of z, we used function composition.
- In (7d) we used the inverse function theorem to both Jacobian terms of the kind $J_h \circ h^{-1}(z,\lambda)$.
- In (7e) we use a property of determinant of invertible matrices, namely, $\det A^{-1} = \frac{1}{\det A}$.
- In (7f) the absolute determinants outside the log cancel and we are left with (7g) where we used the Jacobian of the affine transform.
- Note that $\phi(z)$ does not depend on C and therefore the Jacobian is constant with respect to the standard distribution.

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

  $$\mathbb{E}_{q(x|\lambda)}\left[\log \frac{p(y|x,\theta)p(x|\theta)}{q(x|\lambda)}\right] = \mathbb{E}_{q(x|\lambda)}\left[\log p(y|x,\theta)\right] - \mathrm{KL}\left(q(x|\lambda) \;||\; p(x|\theta)\right) \tag{8}$$

  In case we can compute the KL term analytically, we do not need to included $\left|\det J_{h^{-1}(z,\lambda)}\right|$ in the objective.