Notes on Generalised Reparameterisation Gradient

Wilker Aziz

October 31, 2017

Consider the following function

$$\mathbb{E}_{q(z;\lambda)}\left[\log \frac{g(z)}{q(z;\lambda)}\right] \tag{1}$$

which we mean to maximise by performing stochastic gradient-based optimisation wrt parameters λ. We further assume that Z is a vector valued continuous random variable and that $g(z)$ is differentiable with respect to z.

1 Location-scale q

I first describe the reparameterised gradient when $q(z;\lambda)$ is a location-scale family, that is, $\lambda = \{\mu, C\}$. In such cases, the random variable Z can be represented by transforming samples from a standard distribution $\phi(\epsilon)$ using an affine transformation:

$$\epsilon = h(z;\lambda) = C^{-1}(z-\mu) \tag{2a}$$

$$z = h^{-1}(\epsilon;\lambda) = \mu + C\epsilon \; . \tag{2b}$$

Note that $\phi(\epsilon)$ does not depend on λ which was absorbed in the affine transformation.[1]

For the sake of generality we take z and ϵ to be vector valued. Then we write $J_{h(z;\lambda)}$ to denote the Jacobian matrix of the transformation $h(z;\lambda)$, and $J_{h^{-1}(\epsilon;\lambda)}$ to denote the Jacobian matrix of the inverse transformation.[2] 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 $J_{f^{-1}} \circ f(x) = J^{-1}_{f(x)}$.[3]

For an invertible transformation of random variables, it holds that

$$q(z;\lambda) = \phi(h(z;\lambda))\left|\det J_{h(z;\lambda)}\right| \tag{3}$$

[1] The vector μ is called the *location* and C is a positive definite matrix called the *scale*.

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

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

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

$$q(z;\lambda) = \phi(C^{-1}(z-\mu))\left|\det C^{-1}\right| \tag{4}$$

and

$$\phi(\epsilon) = q(\mu + C\epsilon;\lambda)|\det C| \ . \tag{5}$$

In the following block of equations (7) we will re-express the expectation in Equation (1) in terms of the parameter-free standard density $\phi(\epsilon)$. The derivation relies on several identities, thus we will break it down into small steps. We start by a change of density

$$\int q(z;\lambda)\log\frac{g(z)}{q(z;\lambda)}\mathrm{d}z \tag{6a}$$

$$= \int \phi(\underbrace{h(z;\lambda)}_{\epsilon})\left|\det J_{h(z;\lambda)}\right| \log\frac{g(z)}{\phi(h(z;\lambda))\left|\det J_{h(z;\lambda)}\right|}\mathrm{d}z \tag{6b}$$

where we use the identity in (4) to introduce $\phi(\epsilon)$. Note, however, that the variable of integration is still z and therefore we have expressed every integrand—including $\phi(\epsilon)$—as a function of z. We now proceed to perform a change of variable

$$= \int \phi(\epsilon)\left|\det J_h \circ h^{-1}(\epsilon;\lambda)\right| \log\frac{g(h^{-1}(\epsilon;\lambda))}{\phi(\epsilon)|\det J_h \circ h^{-1}(\epsilon;\lambda)|}\underbrace{\left|\det J_{h^{-1}(\epsilon;\lambda)}\right|\mathrm{d}\epsilon}_{\mathrm{d}z} \tag{6c}$$

which calls for a change of infinitesimal volumes, i.e. $\mathrm{d}z = |\det J_{h^{-1}}(\epsilon;\lambda)|\mathrm{d}\epsilon$, and requires expressing every integrand as a function of ϵ rather than z. Note that, to express the Jacobian $J_{h(z;\lambda)}$ as a function of ϵ, we used function composition. At this point we can use the inverse function theorem

$$= \int \phi(\epsilon)\left|\det J^{-1}_{h^{-1}(\epsilon;\lambda)}\right| \log\frac{g(h^{-1}(\epsilon;\lambda))}{\phi(\epsilon)\left|\det J^{-1}_{h^{-1}(\epsilon;\lambda)}\right|}\left|\det J_{h^{-1}(\epsilon;\lambda)}\right|\mathrm{d}\epsilon \tag{6d}$$

to rewrite both Jacobian terms of the kind $J_h \circ h^{-1}(\epsilon;\lambda)$ as inverse Jacobians. This is convenient because the determinant of invertible matrices is such that $\det A^{-1} = \frac{1}{\det A}$ which we can use to re-arrange the inverse Jacobian terms

$$= \int \phi(\epsilon)\frac{1}{|\det J_{h^{-1}}(\epsilon;\lambda)|} \log\frac{g(h^{-1}(\epsilon;\lambda))|\det J_{h^{-1}}(\epsilon;\lambda)|}{\phi(\epsilon)}|\det J_{h^{-1}}(\epsilon;\lambda)|\mathrm{d}\epsilon \tag{6e}$$

revealing that some of them can be cancelled. We are now left with a simpler expectation wrt $\phi(\epsilon)$

$$= \int \phi(\epsilon)\log\frac{g(h^{-1}(\epsilon;\lambda))|\det J_{h^{-1}}(\epsilon;\lambda)|}{\phi(\epsilon)}\mathrm{d}\epsilon \tag{6f}$$

and we can proceed to solve the Jacobian of the affine transformation

$$= \int \phi(\epsilon)\log\left(g(h^{-1}(\epsilon;\lambda))\left|\det \underbrace{J_{h^{-1}}(\epsilon;\lambda)}_{C}\right|\right)\mathrm{d}\epsilon \underbrace{- \int \phi(\epsilon)\log\phi(\epsilon)\mathrm{d}\epsilon}_{\mathbb{H}[\phi(\epsilon)]} \tag{6g}$$

and to separate out the expected log-denominator (an entropy term). Finally, recall that $\phi(\epsilon)$ does not depend on C and therefore the log-determinant is constant with respect to the standard distribution and can be pushed outside the expectation.

$$= \mathbb{E}_{\phi(\epsilon)}[\log g(h^{-1}(\epsilon;\lambda))] + \log|C| + \mathbb{H}[\phi(\epsilon)] \tag{6h}$$

Note that every expectation in (7h) is taken with respect to $q(\epsilon)$ which does not depend on λ, thus the gradient of (1) wrt λ can be re-expressed as shown in Equation (8).

$$\begin{aligned}
\boldsymbol{\nabla}_\lambda \mathbb{E}_{q(z;\lambda)}\left[\log \frac{g(z)}{q(z;\lambda)}\right] &= \boldsymbol{\nabla}_\lambda \left(\mathbb{E}_{\phi(\epsilon)}[\log g(h^{-1}(\epsilon;\lambda))] + \log|C| + \mathbb{H}[\phi(\epsilon)]\right) && \text{(7a)}\\
&= \mathbb{E}_{\phi(\epsilon)}[\boldsymbol{\nabla}_\lambda \log g(h^{-1}(\epsilon;\lambda))] + \boldsymbol{\nabla}_\lambda \log|C| + \boldsymbol{\nabla}_\lambda \mathbb{H}[\phi(\epsilon)] && \text{(7b)}\\
&= \mathbb{E}_{\phi(\epsilon)}[\underbrace{\boldsymbol{\nabla}_{h^{-1}} \log g(h^{-1}(\epsilon;\lambda))\boldsymbol{\nabla}_\lambda h^{-1}(\epsilon;\lambda)}_{\text{chain rule}}] + \boldsymbol{\nabla}_\lambda \log|C| && \text{(7c)}
\end{aligned}$$

Importantly, note that the first term can be estimated via MC, and that is exactly what automatic differentiation/backprop computes for a given sample, while the second term can be found analytically.