Notes on Generalised Reparameterisation Gradient

Wilker Aziz

October 27, 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.

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 are absorbed in the affine transformation.

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.[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 $J_{f^{-1}} \circ f(x) = J^{-1}_{f(x)}$.[2]

For an invertible transformation of random variables, it holds that

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

and therefore for the transformation in (2) we can write

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

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

and

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

Re-writing the expectation from Equation (1) in terms of the transformed random variable we have

$$
\begin{aligned}
&\int q(z;\lambda) \log \frac{g(z)}{q(z;\lambda)} \mathrm{d}z && \text{(6a)}\\
&= \int \phi(\underbrace{h(z;\lambda)}_{\epsilon})\left|\det J_{h(z;\lambda)}\right| \log \frac{g(z)}{\phi(h(z;\lambda))\left|\det J_{h(z;\lambda)}\right|} \mathrm{d}z && \text{(6b)}\\
&= \int \phi(\epsilon)\left|\det J_h \circ h^{-1}(\epsilon;\lambda)\right| \log \frac{g(h^{-1}(\epsilon;\lambda))}{\phi(\epsilon)|\det J_h \circ h^{-1}(\epsilon;\lambda)|} \left|\det J_{h^{-1}(\epsilon;\lambda)}\right| \mathrm{d}\epsilon && \text{(6c)}\\
&= \int \phi(\epsilon)\left|\det J^{-1}_{h^{-1}(\epsilon;\lambda)}\right| \log \frac{g(h^{-1}(\epsilon;\lambda))}{\phi(\epsilon)\left|\det J^{-1}_{h^{-1}(\epsilon;\lambda)}\right|} \left|\det J_{h^{-1}(\epsilon;\lambda)}\right| \mathrm{d}\epsilon && \text{(6d)}\\
&= \int \phi(\epsilon)\frac{1}{|\det J_{h^{-1}}(\epsilon;\lambda)|} \log \frac{g(h^{-1}(\epsilon;\lambda))|\det J_{h^{-1}}(\epsilon;\lambda)|}{\phi(\epsilon)} |\det J_{h^{-1}}(\epsilon;\lambda)| \mathrm{d}\epsilon && \text{(6e)}\\
&= \int \phi(\epsilon) \log \frac{g(h^{-1}(\epsilon;\lambda))|\det J_{h^{-1}}(\epsilon;\lambda)|}{\phi(\epsilon)} \mathrm{d}\epsilon && \text{(6f)}\\
&= \int \phi(\epsilon) \log \left(g(h^{-1}(\epsilon;\lambda)) \left| \det \underbrace{J_{h^{-1}}(\epsilon;\lambda)}_{C} \right| \right) \mathrm{d}\epsilon - \int \phi(\epsilon) \log \phi(\epsilon) \mathrm{d}\epsilon && \text{(6g)}\\
&= \mathbb{E}_{\phi(\epsilon)}[\log g(h^{-1}(\epsilon,\lambda))] + \log|C| + \mathbb{H}[\phi(\epsilon)] && \text{(6h)}
\end{aligned}
$$

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

A digest of what happened

- In (6b) we applied a change of density.
- In (6c) we applied a change of variable thus expressing every integrand as a function of ϵ rather than z. First, note that this calls for a change of infinitesimal volumes, i.e. $\mathrm{d}z = |\det J_{h^{-1}}(\epsilon,\lambda)|\mathrm{d}\epsilon$. Second, note that, to express the Jacobian $J_{h(z,\lambda)}$ as a function of ϵ, we used function composition.
- In (6d) we used the inverse function theorem to both Jacobian terms of the kind $J_h \circ h^{-1}(\epsilon,\lambda)$.
- In (6e) we use a property of determinant of invertible matrices, namely, $\det A^{-1} = \frac{1}{\det A}$.
- In (6f) the absolute determinants outside the log cancel and we are left with (6g) where we used the Jacobian of the affine transform.

- Note that $\phi(\epsilon)$ does not depend on C and therefore the Jacobian is constant with respect to the standard distribution.