Normalising Flows

Wilker Aziz

November 8, 2018

Abstract

In this note I will present normalising flows and a convenient expression for the ELBO in the general case where both the prior and the posterior approximation are expressed as normalising flows. For that, I first revisit the density of a transformed random variable.

Contents

Notation

Abbreviations

- random variable (rv)
- probability density function (pdf)
- probability mass function (pmf)
- cumulative density function (cdf)

Jacobians For some vector valued function $t : \mathbb{R}^d \to \mathbb{R}^d$, let $\frac{\partial t(a)}{\partial a}$ be the matrix J of partial derivatives where $J_{ij} = \frac{\partial t_i(a)}{\partial a_j}$. This matrix is also called *the Jacobian* and we sometimes denote it $J_t(a)$.

1 Density of transformed variable

For rvs X and Y taking on values in $\mathbb{R}^d$ and an invertible and differentiable transformation h such that $x = h(y)$, it holds that

$$f_X(x) = f_Y(y = h^{-1}(x))\,|\det J_{h^{-1}}(x)| \tag{1}$$

$$f_X(x = t(y)) = f_Y(y)\,|\det J_h(y)|^{-1} \tag{2}$$

however, in this section I will only derive these identities for the scalar case.

Let A and B be rvs taking on values in $\mathbb{R}$. And let $t : \mathbb{R} \to \mathbb{R}$ be an invertible and differentiable transformation such that $a = t(b)$.

As t is invertible, it is either increasing or decreasing in its domain.

Increasing Let's first deal with the case where t is increasing. Then

$$F_A(a) = \Pr\{A \leq a\} \tag{3a}$$

is the definition of A's cdf

$$= \Pr\{t(B) \leq a\} \tag{3b}$$

where we use the fact that $A = t(B)$

$$= \Pr\{B \leq t^{-1}(a)\} \tag{3c}$$

here it is essential that t be increasing and invertible

$$= F_B(t^{-1}(a)) \tag{3d}$$

and finally we realise that we can express F_A in terms of F_B.

Applying the definition of pdfs we obtain

$$f_A(a) = F'_A(a) \tag{4a}$$

$$= \left(F_B(t^{-1}(a))\right)' \tag{4b}$$

where we made use of the result in (3d)

$$= F'_B(t^{-1}(a))\frac{\mathrm{d}t^{-1}(a)}{\mathrm{d}a} \tag{4c}$$

which follows by application of the chain rule of derivatives

$$= f_B(t^{-1}(a))\frac{\mathrm{d}t^{-1}(a)}{\mathrm{d}a} \tag{4d}$$

where again we applied the definition of pdfs.

Decreasing Where t is decreasing things change slightly,

$$F_A(a) = \Pr\{A \le a\} \tag{5a}$$

$$= \Pr\{t(B) \le a\} \tag{5b}$$

$$= \Pr\{B \ge t^{-1}(a)\} \tag{5c}$$

here it is essential that t be decreasing and invertible

$$= 1 - F_B(t^{-1}(a)) \tag{5d}$$

Applying the definition of pdfs we obtain

$$f_A(a) = F_A'(a) \tag{6a}$$

$$= \big(1 - F_B(t^{-1}(a))\big)' \tag{6b}$$

where we made use of the result in (5d)

$$= -F_B'(t^{-1}(a))\frac{\mathrm{d}t^{-1}(a)}{\mathrm{d}a} \tag{6c}$$

$$= -f_B(t^{-1}(a))\frac{\mathrm{d}t^{-1}(a)}{\mathrm{d}a} \tag{6d}$$

A general rule Combining both results (4d) and (6d), we have

$$f_A(a) = f_B(t^{-1}(a))\left|\frac{\mathrm{d}t^{-1}(a)}{\mathrm{d}a}\right| \tag{7}$$

which can also be written

$$f_A(a = t(b)) = f_B(b)\left|\frac{\mathrm{d}t(b)}{\mathrm{d}a}\right|^{-1} . \tag{8}$$

2 Change of infinitesimal volume

I will only discuss the scalar case. Let $z \in \mathbb{R}$ and $z = t(h)$ for $h \in \mathbb{R}$, then

$$\frac{\mathrm{d}a}{\mathrm{d}b} = \frac{\mathrm{d}t(b)}{\mathrm{d}b} \tag{9}$$

and therefore

$$\mathrm{d}a = \frac{\mathrm{d}t(b)}{\mathrm{d}b}\mathrm{d}b \tag{10}$$

where $\frac{\mathrm{d}t(b)}{\mathrm{d}b}$ is the determinant of the Jacobian of the scalar transformation. All we really need is to scale the volume (the direction does not matter), thus we use the absolute value

$$\mathrm{d}a = \left|\frac{\mathrm{d}t(b)}{\mathrm{d}b}\right|\mathrm{d}b . \tag{11}$$

I know I am being very informal here, but this "rule" should help you throughout. Then in high dimensions it becomes

$$\mathrm{d}x = |\det J_t(y)|\,\mathrm{d}y . \tag{12}$$

3 Normalising flows

We can define a rich density by transforming a random variable sampled from a base distribution, for which we have a closed-form pdf, using an invertible transformation. We can then use rule (1) to obtain a density for the transformed variable. This is what we call a *normalising flow*, where we call *flow* the transformation itself.

A normalising flow is nothing but application of rule (1), which I repeat here for a density $p_Z(z)$ where $z = s(\gamma)$ for some $\gamma \sim \tau(\cdot)$, and s is invertible.[1]

$$p(z) = \tau(\gamma = s^{-1}(z)) \left|\det J_{s^{-1}}(z)\right| \tag{13a}$$

$$p(z = s(\gamma)) = \tau(\gamma) \left|\det J_s(\gamma)\right|^{-1} \tag{13b}$$

3.1 Variational inference with normalising flows

In this section I will assume a generative model

$$p(x, z) = p(z)p(x|z) \tag{14}$$

for which we assume an intractable posterior $p(z|x)$. We then approximate this intractable posterior by a proxy density $q(z)$ in the framework of variational inference.

We let our prior be a normalising flow (13) and similarly design a posterior normalising flow:

$$q(z) = \pi(\epsilon = t^{-1}(z)) \left|\det J_{t^{-1}}(z)\right| \tag{15a}$$

$$q(z = t(\epsilon)) = \pi(\epsilon) \left|\det J_t(\epsilon)\right|^{-1} \quad . \tag{15b}$$

In variational inference we optimise a variational lowerbound knowns as ELBO, here we obtain an expression for the ELBO as expectations wrt to the base distribution of the posterior flow. We start from the definition of the ELBO

$$\mathbb{E}_{q(z)}\left[\log \frac{p(x, z)}{q(z)}\right] \tag{16a}$$

$$= \int q(z) \log \frac{p(x|z)p(z)}{q(z)} \mathrm{d}z \tag{16b}$$

where we have simply factorised the joint distribution

$$= \int \pi(\epsilon = t^{-1}(z)) \left|\det J_{t^{-1}}(z)\right| \log \frac{p(x|z)p(z)}{\pi(\epsilon = t^{-1}(z)) \left|\det J_{t^{-1}}(z)\right|} \mathrm{d}z \tag{16c}$$

[1] In this section I will drop the subscript to reduce clutter, and to avoid confusion I will use different letters for different densities, for example, p for p_Z and q for q_Z.

here we expressed the posterior flow in terms of its base distribution, but note we have not yet changed the variable of integration

$$= \int \pi(\epsilon) \left|\det J_t(\epsilon)\right|^{-1} \log \frac{p(x|z = t(\epsilon))p(z = t(\epsilon))}{\pi(\epsilon = t^{-1}(z)) \left|\det J_t(\epsilon)\right|^{-1}} \left|\det J_t(\epsilon)\right| \mathrm{d}\epsilon \tag{16d}$$

here we performed a change of variable expressing the integral wrt ϵ

$$= \int \pi(\epsilon) \log \frac{p(x|z = t(\epsilon))p(z = t(\epsilon))}{\pi(\epsilon) \left|\det J_t(\epsilon)\right|^{-1}} \mathrm{d}\epsilon \tag{16e}$$

note that some of the Jacobians cancel out

$$= \int \pi(\epsilon) \log \frac{p(x|z = t(\epsilon))\tau(\gamma = s^{-1}(t(\epsilon)) \left|\det J_{s^{-1}}(t(\epsilon))\right|}{\pi(\epsilon) \left|\det J_t(\epsilon)\right|^{-1}} \mathrm{d}\epsilon \tag{16f}$$

we now expressed the prior flow in terms of its base distribution

$$= \mathbb{E}_{\pi(\epsilon)} \left[\log p(x|z = t(\epsilon))\right] + \mathbb{H}(\pi(\epsilon)) \tag{16g}$$

$$+ \mathbb{E}_{\pi(\epsilon)} \left[\log \tau(\gamma = s^{-1}(t(\epsilon)))\right] + \mathbb{E}_{\pi(\epsilon)} \left[\log \frac{\left|\det J_{s^{-1}}(t(\epsilon))\right|}{\left|\det J_t(\epsilon)\right|^{-1}}\right] \tag{16h}$$

and after separating a few terms out we identify a convenient expression where some terms may even be available in closed-form (e.g. the entropy of the posterior flow's base distribution). Note that to estimate (16f) we need to assess z, sampled from the posterior normalising flow, under the prior normalising flow. As we have two different flows, and as we are sampling from the posterior base distribution, we need to first find the posterior sample $z = t(\epsilon)$, then find the prior sample $\gamma = s^{-1}(z) = s^{-1}(t(\epsilon))$, finally with that γ, we can compute the determinant of the Jacobian of the inverse prior flow s^{-1} and evaluate it at the prior sample γ.