Supplementary Materials: Pathwise Derivatives Beyond the Reparameterization Trick

Martin Jankowiak [*1] **Fritz Obermeyer** [*1]

1. The Univariate Case

For completeness we show explicitly that the formula

$$\frac{dz}{d\theta} = -\frac{\frac{\partial F_{\boldsymbol\theta}}{\partial\theta}(z)}{q_{\boldsymbol\theta}(z)} \tag{1}$$

yields the correct gradient. Without loss of generality we assume that $f(z)$ has no explicit dependence on θ. Substituting Eqn. 1 for $\frac{dz}{d\theta}$ we have

$$\begin{aligned}\mathbb{E}_{q_{\boldsymbol\theta}(z)}\left[\frac{\partial f}{\partial z}\frac{\partial z}{\partial\theta}\right] &= -\int_{-\infty}^{\infty}\frac{q_{\boldsymbol\theta}(z)}{q_{\boldsymbol\theta}(z)}\frac{\partial f}{\partial z}\int_{-\infty}^{z}\frac{\partial q_{\boldsymbol\theta}(z')}{\partial\theta}dz'dz \\ &= -\int_{-\infty}^{\infty}\frac{\partial q_{\boldsymbol\theta}(z')}{\partial\theta}\int_{z'}^{\infty}\frac{\partial f}{\partial z}dzdz' \\ &= -\int_{-\infty}^{\infty}\frac{\partial q_{\boldsymbol\theta}(z')}{\partial\theta}\left(-f(z')\right)dz' \\ &= \frac{d}{d\theta}E_{q_{\boldsymbol\theta}(z)}[f(z)]\end{aligned} \tag{2}$$

In the second line we changed the order of integration and in the third we appealed to the fundamental theorem of calculus, assuming that $f(z)$ is sufficiently regular that we can drop the boundary term at infinity.

Note that Eqn. 1 is the unique solution $v = \frac{dz}{d\theta}$ to the one-dimensional version of the transport equation that satisfies the boundary condition $\lim_{z\to\infty} q_{\boldsymbol\theta} v = 0$:

$$\frac{\partial q_{\boldsymbol\theta}}{\partial\theta} + \frac{\partial}{\partial z}\left(q_{\boldsymbol\theta} v\right) = 0 \tag{3}$$

1.1. Example: Truncated Unit Normal

We consider an illustrative case where Eqn. 1 can be computed in closed form. For simplicity we consider the unit Normal distribution truncated[1] to the interval $[0, \kappa]$ with κ as the only free parameter. A simple computation yields

$$\frac{dz}{d\kappa} = e^{\frac{1}{2}(z^2-\kappa^2)}\frac{\mathrm{erf}(\frac{z}{\sqrt{2}})}{\mathrm{erf}(\frac{\kappa}{\sqrt{2}})} \tag{4}$$

First, notice that for $z = \kappa$ we have $\frac{dz}{d\kappa} = 1$, which is what we would expect, since $u = 1$ is mapped to the rightmost edge of the interval at $z = \kappa$, i.e. $F_\kappa^{-1}(1) = \kappa$. Similarly we have $\frac{dz}{d\kappa} = 0$ for $z = 0$. For $z \in (0, \kappa)$ the derivative $\frac{dz}{d\kappa}$ interpolates smoothly between 0 and 1. This makes sense, since for a fixed value of u as we get further into the tails of the distribution, nudging κ to the right has a correspondingly larger effect on $z = F_\kappa^{-1}(u)$, while it has a correspondingly smaller effect for u in the bulk of the distribution.

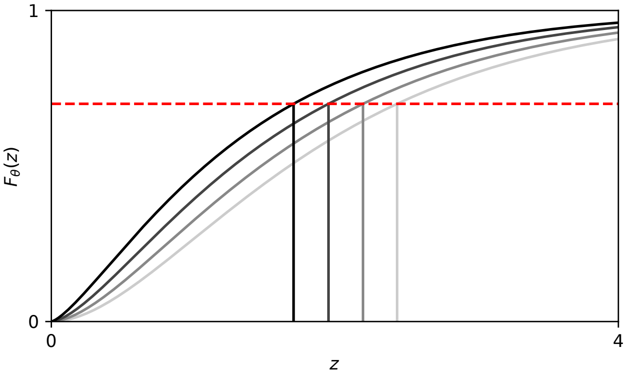

Figure 1. We illustrate how the pathwise derivative is obtained from the CDF in the univariate case. The black curves depict the CDF of the Gamma distribution with $\beta = 1$ and α varying between 1.4 and 2.0. The red line corresponds to a fixed quantile u. As we vary α the point z where the CDF intersects the red line varies. The rate of this variation is precisely the derivative $\frac{dz}{d\alpha}$.

[*] Equal contribution [1] Uber AI Labs, San Francisco, USA. Correspondence to: <jankowiak@uber.com>, <fritzo@uber.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

[1] As one would expect, Eqn. 1 yields the standard reparameterized gradient in the case of an non-truncated Normal distribution. Also note that the truncated unit normal is amenable to the reparameterization trick provided that one can compute the inverse error function erf^{-1}.

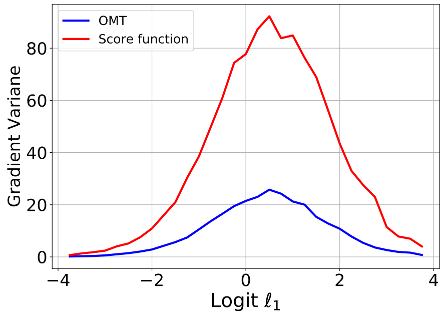

Figure 2. We compare the OMT gradient to the score function gradient for the test function $f(z) = z^4$ where $q_{\boldsymbol\theta}(z)$ is a mixture with two components. Depicted is the variance of the gradient w.r.t. the logit ℓ_1 that governs the mixture probability of the first component. The logit of the second component is fixed to be zero.

1.2. Example: Univariate Mixture Distributions

Consider a mixture of univariate distributions:

$$q_{\boldsymbol\theta}(z) = \sum_{k=1}^{K} \pi_k q_{\theta_k}(z) \tag{5}$$

If we have analytic control over the individual CDFs (or know how to approximate them and their derivatives w.r.t. the parameters) then we can immediately appeal to Eqn. 1. Concretely for derivatives w.r.t. the parameters of each component distribution we have:

$$\frac{\partial z}{\partial \theta_i} = -\frac{\pi_i \frac{\partial F_{\theta_i}}{\partial \theta_i}(z)}{q_{\boldsymbol\theta}(z)} \tag{6}$$

from which we can get, for example

$$\frac{\partial z}{\partial \mu_i} = \frac{\pi_i q_{\mu_i,\sigma_i}(z)}{q_{\boldsymbol\theta}(z)} \tag{7}$$

for a mixture of univariate Normal distributions.

In Fig. 2 we demonstrate that the OMT gradient for a mixture of univariate Normal distributions can have much lower variance than the corresponding score function gradient. Here the mixture has two components with $\boldsymbol\mu = (0, 1)$ and $\boldsymbol\sigma = (1, 1)$. Note that using the reparameterization trick in this setting would be impractical.

2. The Multivariate Case

Suppose we are given a velocity field that satisfies the transport equation:

$$\frac{\partial}{\partial\theta} q_{\boldsymbol\theta} + \nabla_{\boldsymbol z} \cdot \left(q_{\boldsymbol\theta} \boldsymbol v^\theta\right) = 0 \tag{8}$$

Then, as discussed in the main text, we can form the gradient estimator

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{q_{\boldsymbol\theta}(\boldsymbol z)}\left[\boldsymbol v^\theta \cdot \nabla_{\boldsymbol z} f\right] \tag{9}$$

That this gradient estimator is unbiased follows directly from the transport equation and divergence theorem:

$$\nabla_\theta \mathcal{L} = \int d\boldsymbol z \frac{\partial q_{\boldsymbol\theta}(\boldsymbol z)}{\partial \theta} f(\boldsymbol z) = -\int d\boldsymbol z \nabla_{\boldsymbol z} \cdot \left(q_{\boldsymbol\theta} \boldsymbol v^\theta\right) f(\boldsymbol z) = \int d\boldsymbol z q_{\boldsymbol\theta}(\boldsymbol z) \nabla_{\boldsymbol z} f \cdot \boldsymbol v^\theta = \mathbb{E}_{q_{\boldsymbol\theta}(\boldsymbol z)}\left[\nabla_{\boldsymbol z} f \cdot \boldsymbol v^\theta\right] \tag{10}$$

where we appeal to the identity

$$\int_V f \nabla_{\boldsymbol z} \cdot (q_{\boldsymbol\theta} \boldsymbol v^\theta)\, \mathrm{d}V = -\int_V \nabla_{\boldsymbol z} f \cdot (q_{\boldsymbol\theta} \boldsymbol v^\theta)\, \mathrm{d}V + \oint_S (q_{\boldsymbol\theta} f \boldsymbol v^\theta) \cdot \hat{\mathbf n}\, \mathrm{d}S \tag{11}$$

and assume that $q_{\boldsymbol\theta} f \boldsymbol v^\theta$ is sufficiently well-behaved that we can drop the surface integral. This is just the multivariate generalization of the derivation in the previous section.

3. Multivariate Normal

3.1. Whitened Coordinates

First we take a look at gradient estimators in whitened coordinates $\tilde{\boldsymbol z} = L^{-1}\boldsymbol z$. The reparameterization trick ansatz for the velocity field can be obtained by transforming the solution in Eqn. 26 (which is also given in the main text) to the new coordinates:

$$\tilde v_i \equiv \frac{\partial \tilde z_i}{\partial L_{ab}} = L^{-1}_{ia} \tilde z_b \tag{12}$$

Note that the transport equation for the multivariate distribution can be written in the form

$$\frac{\partial}{\partial L_{ab}} \log q + \nabla \cdot \tilde{\boldsymbol v} + \tilde{\boldsymbol v} \cdot \nabla \log q = 0 \tag{13}$$

The homogenous equation (i.e. the transport equation without the source term $\frac{\partial \log q}{\partial L_{ab}}$) is then given by

$$\nabla \cdot \tilde{\boldsymbol v} = \tilde{\boldsymbol v} \cdot \tilde{\boldsymbol z} \tag{14}$$

In these coordinates it is evident that infinitesimal rotations, i.e. vector fields of the form

$$\tilde w_i = (A \tilde{\boldsymbol z})_i \qquad \text{with} \qquad A_{ij} = -A_{ji} \tag{15}$$

satisfy[2] the homogenous equation, since

$$\nabla \cdot \tilde{\boldsymbol w} = \mathrm{Tr}\ A = 0 = \sum_{ij} \tilde z_i A_{ij} \tilde z_j = \tilde{\boldsymbol w} \cdot \tilde{\boldsymbol z} \tag{16}$$

[2] These are in fact not the only solutions; in addition there are non-linear solutions.

Finally, if we make the specific choice

$$A_{ij} = \frac{1}{2}\left(\delta_{ib}L^{-1}_{ja} - \delta_{jb}L^{-1}_{ia}\right) \tag{17}$$

we find that $\tilde{v}_i + \tilde{w}_i$ (which automatically satisfies the transport equation) and which is given by

$$\tilde{v}_i + \tilde{w}_i \equiv \left(\frac{\partial \tilde{z}_i}{\partial L_{ab}}\right)^{\mathrm{OMT}} = \frac{1}{2}\left(L^{-1}_{ia}\tilde{z}_b + \delta_{ib}\sum_k L^{-1}_{ka}\tilde{z}_k\right)$$

satisfies the symmetry condition

$$\frac{\partial}{\partial \tilde{z}_j}\left(\frac{\partial \tilde{z}_i}{\partial L_{ab}}\right)^{\mathrm{OMT}} = \frac{\partial}{\partial \tilde{z}_i}\left(\frac{\partial \tilde{z}_j}{\partial L_{ab}}\right)^{\mathrm{OMT}} \tag{18}$$

since

$$\frac{\partial}{\partial \tilde{z}_j}\left(\frac{\partial \tilde{z}_i}{\partial L_{ab}}\right)^{\mathrm{OMT}} = \frac{1}{2}\left(L^{-1}_{ia}\delta_{jb} + L^{-1}_{ja}\delta_{ib}\right) \tag{19}$$

which is symmetric in i and j. This implies that the velocity field can be specified as the gradient of a scalar field (this is generally true for the OMT solution), i.e.

$$\left(\frac{\partial \tilde{z}_i}{\partial L_{ab}}\right)^{\mathrm{OMT}} = \frac{\partial}{\partial \tilde{z}_i}\tilde{T}^{ab}(\tilde{\boldsymbol{z}}) \tag{20}$$

for some $\tilde{T}^{ab}(\tilde{\boldsymbol{z}})$, which is evidently given by[3]

$$\tilde{T}^{ab}(\tilde{\boldsymbol{z}}) = \frac{1}{2}(L^{-\mathrm{T}}\tilde{z})_a\tilde{z}_b \tag{21}$$

Note, however, that this is not the OMT solution we care about: it minimizes a *different* kinetic energy functional to the one we care about (namely it minimizes the kinetic energy functional in whitened coordinates and not in natural coordinates).

We now explicitly show that solutions of the transport equation that are modified by the addition of an infinitesimal rotation (as in Eqn. 18) still yield valid gradient estimators. Consider a test statistic $f(\tilde{\boldsymbol{z}})$ that is a monomial in $\tilde{\boldsymbol{z}}$:

$$f(\tilde{\boldsymbol{z}}) = \kappa\prod_{i=1}^{n}\tilde{z}_i^{n_i} \tag{22}$$

It is enough to show that the following expectation vanishes:[4]

$$\mathbb{E}_{q_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}})}\left[\sum_{ij}\frac{\partial f}{\partial \tilde{z}_i}A_{ij}\tilde{z}_j\right] \tag{23}$$

where A_{ij} is an antisymmetric matrix. The sum in Eqn. 23 splits up into a sum of paired terms of the form

$$\mathbb{E}_{q_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}})}\left[A_{ij}\left(\frac{\partial f}{\partial \tilde{z}_i}\tilde{z}_j - \frac{\partial f}{\partial \tilde{z}_j}\tilde{z}_i\right)\right] \tag{24}$$

We can easily show that each of these paired terms has zero expectation. First note that the expectation is zero if either of i or j is even (since $\mathbb{E}_{q_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}})}\left[\tilde{z}_l^{2k-1}\right] = 0$). If both i and j are odd we get (using $\mathbb{E}_{q_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}})}\left[\tilde{z}_l^{2k}\right] = (2k-1)!!$, where !! is the double factorial)

$$\kappa A_{ij}\left[n_i(n_i-2)!!n_j!! - n_j(n_j-2)!!n_i!!\right] = 0 \tag{25}$$

Thus, solutions of the transport equation that are modified by the addition of an infinitesimal rotation still yield the same gradient $\nabla_{L_{ab}}\mathbb{E}_{q_{\boldsymbol{\theta}}(\tilde{\boldsymbol{z}})}\left[f(\tilde{\boldsymbol{z}})\right]$ in expectation.

3.2. Natural Coordinates

We first show that the velocity field $\boldsymbol{v}^{\mathrm{RT}}$ that follows from the reparameterization trick satisfies the transport equation in the (given) coordinates $\boldsymbol{z}$, where we have

$$v_i^{\mathrm{RT}} \equiv \frac{\partial z_i}{\partial L_{ab}} = \delta_{ia}(L^{-1}\boldsymbol{z})_b \tag{26}$$

We have that

$$\begin{aligned}\frac{\partial \log q}{\partial L_{ab}} &= \frac{\partial}{\partial L_{ab}}\left(-\log\det L - \frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{z}\right)\\ &= -L^{-1}_{ba} + \left(\Sigma^{-1}\boldsymbol{z}\right)_a\left(L^{-1}\boldsymbol{z}\right)_b\end{aligned} \tag{27}$$

and

$$\nabla\cdot\boldsymbol{v}^{\mathrm{RT}} = L^{-1}_{ba} \tag{28}$$

and

$$\boldsymbol{v}^{\mathrm{RT}}\cdot\nabla\log q = -\boldsymbol{v}^{\mathrm{RT}}\cdot\left(\Sigma^{-1}\boldsymbol{z}\right) = -\left(\Sigma^{-1}\boldsymbol{z}\right)_a\left(L^{-1}\boldsymbol{z}\right)_b$$

Thus, the terms cancel term by term and the transport equation is satisfied.

What about the OMT gradient in the natural (given) coordinates $\boldsymbol{z}$? To proceed we represent $\boldsymbol{v}$ as a linear vector field with symmetric and antisymmetric parts. Imposing the OMT condition determines the antisymmetric part. Imposing the transport equation determines the symmetric part. We find that

$$v_i^{\mathrm{OMT}} = \frac{1}{2}\left(\delta_{ia}(L^{-1}\boldsymbol{z})_b + z_aL^{-1}_{bi}\right) + (S^{ab}\boldsymbol{z})_i \tag{29}$$

where S^{ab} is the unique symmetric matrix that satisfies the equation

$$\Sigma^{-1}S^{ab} + S^{ab}\Sigma^{-1} = \Xi^{ab} \quad \text{with} \quad \Xi^{ab} \equiv \xi^{ab} + (\xi^{ab})^{\mathrm{T}}$$

where we define

$$\xi^{ab}_{ij} = \tfrac{1}{2}\left(L^{-1}_{bi}\Sigma^{-1}_{aj} - \delta_{ai}(L^{-1}\Sigma^{-1})_{bj}\right) \tag{30}$$

To explicitly solve Eqn. 30 for S^{ab} we use SVD to write

$$\Sigma^{-1} = UDU^{\mathrm{T}} \qquad \text{and} \qquad \tilde{\Xi}^{ab} = U^{\mathrm{T}}\Xi^{ab}U \tag{31}$$

[3] Up to an unspecified additive constant.

[4] Note that we can thus think of this term as a control variate.

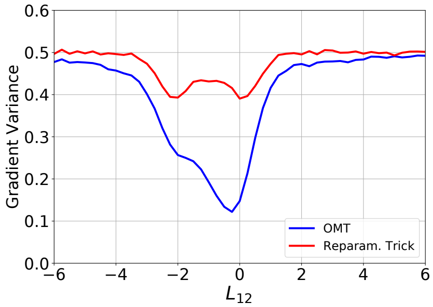

Figure 3. We compare the OMT gradient to the gradient from the reparameterization trick for a bivariate Normal distribution and the test function $f_{\boldsymbol{\theta}}(\boldsymbol{z}) = \cos \boldsymbol{\omega} \cdot \boldsymbol{z}$ with $\boldsymbol{\omega} = (1, 1)$. The Cholesky factor $\boldsymbol{L}$ has diagonal elements $(1, 1)$ and off-diagonal element L_{21}. The gradient is with respect to L_{21}. The variance for the OMT gradient is everywhere lower than for the reparameterization trick gradient.

where D and U are diagonal and orthogonal matrices, respectively. Then we have that

$$S^{ab} = U\left(\tilde{\Xi}^{ab} \div (D \otimes \mathbb{1} + \mathbb{1} \otimes D)\right) U^{\mathrm{T}} \tag{32}$$

where $\div$ represents elementwise division and $\otimes$ is the outer product. Note that a naive implementation of a gradient estimator based on Eqn. 29 would explicitly construct ξ_{ij}^{ab}, which has size quartic in the dimension. A more efficient implementation will instead make use of ξ_{ij}^{ab}'s structure as a sum of products and never explicitly constructs ξ_{ij}^{ab}.[5]

3.3. Bivariate Normal distribution

In Fig. 3 we compare the performance of our OMT gradient for a bivariate Normal distribution to the reparameterization trick gradient estimator. We use a test function $f_{\boldsymbol{\theta}}(\boldsymbol{z})$ for which we can compute the gradient exactly. We see that the OMT gradient estimator performs favorably over the entire range of parameters considered.

4. Gradient Variance for Linear Test Functions

We use the following example to give more intuition for when we expect OMT gradients for the multivariate Normal distribution to be lower variance than RT gradients. Let $q_{\boldsymbol{\theta}}(\boldsymbol{z})$ be the unit normal distribution in D dimensions. Consider the test function

$$f(\boldsymbol{z}) = \sum_{i=1}^{D} \kappa_i z_i \qquad \mathcal{L} = \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})}\left[f(\boldsymbol{z})\right] \tag{33}$$

and the derivative w.r.t. the off-diagonal elements of the Cholesky factor L. A simple computation yields the total variance of the RT estimator:

$$\sum_{a>b} \mathrm{Var}\left(\frac{\partial \mathcal{L}}{\partial \mathrm{L_{ab}}}\right) = \sum_{\mathrm{a>b}} \kappa_{\mathrm{a}}^2 \tag{34}$$

Similarly for the OMT estimator we find

$$\sum_{a>b} \mathrm{Var}\left(\frac{\partial \mathcal{L}}{\partial \mathrm{L_{ab}}}\right) = \frac{1}{4} \sum_{\mathrm{a>b}} \left(\kappa_{\mathrm{a}}^2 + \kappa_{\mathrm{b}}^2\right) \tag{35}$$

So if we draw the parameters κ_i from a generic prior we expect the variance of the OMT estimator to be about half of that of the RT estimator. Concretely, if $\kappa_i \sim \mathcal{N}(0, 1)$ then the variance of the OMT estimator will be exactly half that of the RT estimator in expectation. While this computation is for a very specific case—a linear test function and a unit normal $q_{\boldsymbol{\theta}}(\boldsymbol{z})$—we find that this magnitude of variance reduction is typical.

5. The Lugannani-Rice Approximation

Saddlepoint approximation methods take advantage of cumulant generating functions (CGFs) to construct (often very accurate) approximations to probability density functions in situations where full analytic control is intractable.[6] These methods are also directly applicable to CDFs, where a particularly useful approximation—often used by statisticians to estimate various tail probabilities—has been developed by Lugannani and Rice (Lugannani & Rice, 1980). This approximation—after additional differentiation w.r.t. the parameters of the distribution $q_{\boldsymbol{\theta}}(z)$—forms the basis of our approximate formulas for pathwise gradients for the Gamma, Beta and Dirichlet distributions in regions of (z, θ) where the (marginal) density is approximately gaussian. As we will see these approximations attain high accuracy.

For completeness we briefly describe the Lugannani-Rice approximation. It is given by:

$$F(z) \approx \begin{cases} \Phi(\hat{w}) + \phi(\hat{w})(1/\hat{w} - 1/\hat{u}) & \text{if } z \neq \mu \\ \frac{1}{2} + \frac{K'''(0)}{6\sqrt{2\pi}K''(0)^{3/2}} & \text{if } z = \mu \end{cases} \tag{36}$$

where

$$\hat{w} = \mathrm{sgn}(\hat{s})\sqrt{2\{\hat{s}z - K(\hat{s})\}} \qquad \hat{u} = \hat{s}\sqrt{K''(\hat{s})} \tag{37}$$

and where $\hat{w}$ and $\hat{u}$ are functions of z and the saddlepoint $\hat{s}$, with the saddlepoint defined implicitly by the equation $K'(\hat{s}) = z$. Here $K(s) = \log \mathbb{E}_{q_{\boldsymbol{\theta}}(z)}\left[\exp(sz)\right]$ is the CGF of $q_{\boldsymbol{\theta}}(z)$, μ is the mean of $q_{\boldsymbol{\theta}}(z)$, and $\Phi(\cdot)$ and $\phi(\cdot)$ are the CDFs and probability densities of the unit normal distribution. Note that Eqn. 36 appears to have a singularity at

[5] Our implementation can be found here: https://github.com/uber/pyro/blob/0.2.1/pyro/distributions/omt_mvn.py

[6] We refer the reader to (Butler, 2007) for an overview.

$z = \mu$; it can be shown, however, that Eqn. 36 is in fact smooth at $z = \mu$. Nevertheless, in our numerical recipes we will need to take care to avoid numerical instabilities near $z = \mu$ that result from finite numerical precision.

6. Gamma Distribution

Our numerical recipe for $\frac{dz}{d\alpha}$ for the standard Gamma distribution with $\beta = 1$ divides (z, α) space into three regions. If $z < 0.8$ we use the Taylor series expansion given in the main text. If $\alpha > 8$ we use the following set of expressions derived from the Lugannani-Rice approximation. Away from the singularity, for $z \gtrless \alpha \pm \delta \cdot \alpha$, we use:

$$\frac{dz}{d\alpha} = \frac{\sqrt{\frac{2}{\alpha}\frac{\alpha+z}{(\alpha-z)^2}} + \log\frac{z}{\alpha}\left(\frac{\sqrt{8\alpha}}{z-\alpha} \pm (z-\alpha-\alpha\log\frac{z}{\alpha})^{-\frac{3}{2}}\right)}{\sqrt{8\alpha}/(z\mathcal{S}_\alpha)} \tag{38}$$

where

$$\mathcal{S}_\alpha \equiv 1 + \frac{1}{12\alpha} + \frac{1}{288\alpha^2}$$

Near the singularity, i.e. for $|z - \alpha| \leq \delta \cdot \alpha$, we use:

$$\frac{dz}{d\alpha} = \frac{1440\alpha^3 + 6\alpha z(53-120z) - 65z^2 + \alpha^2(107+3600z)}{1244160\alpha^5/(1+24\alpha+288\alpha^2)} \tag{39}$$

Note that Eqn. 39 is derived from Eqn. 38 by a Taylor expansion in powers of $(z-\alpha)$. We set $\delta = 0.1$, which is chosen to balance use of Eqn. 38 (which is more accurate) and Eqn. 39 (which is more numerically stable for $z \approx \alpha$). Finally, in the remaining region ($z > 0.8$ and $\alpha < 8$) we use a bivariate rational polynomial approximation $f(z, \alpha) = \exp\left(\frac{p(z,\alpha)}{q(z,\alpha)}\right)$ where p, q are polynomials in the coordinates $\log(z/\alpha)$ and $\log(\alpha)$, with terms up to order 2 in $\log(z/\alpha)$ and order 3 in $\log(\alpha)$. We fit the rational approximation using least squares on 15696 random (z, α) pairs with α sampled log uniformly between 0.00001 and 10, and z sampled conditioned on α. Our complete approximation for $\frac{dz}{d\alpha}$ is unit tested to have relative accuracy of 0.0005 on a wide range of inputs.

7. Beta Distribution

The CDF of the Beta distribution is given by

$$F_{\alpha,\beta}(z) = \frac{B(z;\alpha,\beta)}{B(\alpha,\beta)} \tag{40}$$

where $B(z;\alpha,\beta)$ and $B(\alpha,\beta)$ are the incomplete beta function and beta function, respectively. Our numerical recipe for computing $\frac{dz}{d\alpha}$ and $\frac{dz}{d\beta}$ for the Beta distribution divides (z, α, β) space into three sets of regions. First suppose that $z \ll 1$. Then just like for the Gamma distribution, we can compute a Taylor series of $B(z;\alpha,\beta)$ in powers of z

$$B(z;\alpha,\beta) = z^\alpha\left(\frac{1}{\alpha} + \frac{1-\beta}{1+\alpha}z + \frac{1-\frac{3\beta}{2}+\frac{\beta^2}{2}}{2+\alpha}z^2 + ...\right) \tag{41}$$

that can readily be differentiated w.r.t. either α or β. Combined with the derivatives of the beta function,

$$\begin{aligned}\frac{d}{d\alpha}B(\alpha,\beta) &= B(\alpha,\beta)\left(\psi(\alpha) - \psi(\alpha+\beta)\right)\\ \frac{d}{d\beta}B(\alpha,\beta) &= B(\alpha,\beta)\left(\psi(\beta) - \psi(\alpha+\beta)\right)\end{aligned} \tag{42}$$

this gives a complete recipe for approximating $\frac{dz}{d\alpha}$ and $\frac{dz}{d\beta}$ for small z.[7] By appealing to the symmetry of the Beta distribution

$$\mathrm{Beta(z|}\alpha,\beta) = \mathrm{Beta(1-z|}\beta,\alpha) \tag{43}$$

we immediately gain approximations to $\frac{dz}{d\alpha}$ and $\frac{dz}{d\beta}$ for $1 - z \ll 1$. It remains to specify when these various approximations are applicable. Let us define $\xi = z(1-z)(\alpha+\beta)$. Empirically we find that these approximations are accurate for $\frac{dz}{d\alpha}$ if

1. $z \leq 0.5$ and $\xi < 2.5$; or
2. $z \geq 0.5$ and $\xi < 0.75$

with the conditions flipped for $\frac{dz}{d\beta}$. Depending on the precise region, we use 8 to 10 terms in the Taylor series.

Next we describe the set of approximations we derived from the Lugannani-Rice approximation and that we find to be accurate for $\alpha > 6$ and $\beta > 6$. By Eqn. 43 it is sufficient to describe our approximation for $\frac{dz}{d\alpha}$. First define $\sigma = \frac{\sqrt{\alpha\beta}}{(\alpha+\beta)\sqrt{\alpha+\beta+1}}$, the standard deviation of the Beta distribution. Then away from the singularity, for $z \gtrless \frac{\alpha}{\alpha+\beta} \pm \epsilon \cdot \sigma$, we use:

$$\frac{dz}{d\alpha} = \frac{z(1-z)\left(\mathcal{A} + \log\frac{\alpha}{z(\alpha+\beta)}\mathcal{B}_\pm\right)}{\sqrt{\frac{2\alpha\beta}{\alpha+\beta}\frac{S_{\alpha\beta}}{S_\alpha S_\beta}}} \tag{44}$$

with

$$\mathcal{A} = \frac{\beta(2\alpha^2(1-z) + \alpha\beta(1-z) + \beta^2 z)}{\sqrt{2\alpha\beta}(\alpha+\beta)^{3/2}(\alpha(1-z) - \beta z)^2}$$

and

$$\mathcal{B}_\pm = \frac{\sqrt{\frac{2\alpha\beta}{\alpha+\beta}}}{\alpha(1-z)-\beta z} \pm \frac{1}{2}\left(\alpha\log\frac{\alpha}{(\alpha+\beta)(1-z)} + \beta\log\frac{\beta}{(\alpha+\beta)z}\right)^{-3/2}$$

Near the singularity, i.e. for $|z - \frac{\alpha}{\alpha+\beta}| \leq \epsilon \cdot \sigma$, we use:

$$\frac{dz}{d\alpha} = \frac{(12\alpha+1)(12\beta+1)(\mathcal{H}+\mathcal{I}+\mathcal{J}+\mathcal{K})}{12960\alpha^3\beta^2(\alpha+\beta)^2(12\alpha+12\beta+1)} \tag{45}$$

[7] Here $\psi(\cdot)$ is the digamma function, which is available in most advanced tensor libraries.

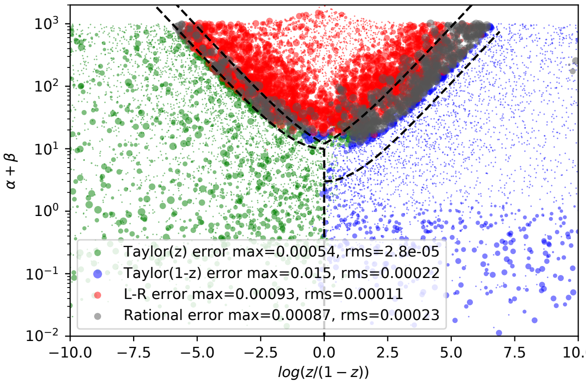

Figure 4. Relative error of our four approximations for $\frac{dz}{d\alpha}$ for the Beta distribution in their respective regions. Note that the region boundaries are in the three-dimensional z, α, β space, so the upper boundaries are only cross-sections.

with

$$\begin{aligned}
\mathcal{H} &= 8\alpha^4(135\beta - 11)(1 - z)\\
\mathcal{I} &= \alpha^3\beta(453 - 455z + 1620\beta(1 - z))\\
\mathcal{J} &= 3\alpha^2\beta^2(180\beta - 90z + 59)\\
\mathcal{K} &= \alpha\beta^3(20z(27\beta + 16) + 43) + 47\beta^4 z
\end{aligned}$$

We set $\epsilon = 0.1$, which is chosen to balance numerical accuracy and numerical stability (just as in the case of the Gamma distribution).

Finally, in the remaining region we use a rational multivariate polynomial approximation

$$f(z, \alpha, \beta) = \frac{p(z, \alpha, \beta)}{q(z, \alpha, \beta)} \frac{z(1 - z)}{\beta} \left(\psi(\alpha + \beta) - \psi(\alpha)\right)$$

where p, q are polynomials in the three coordinates $\log(z)$, $\log(\alpha/z)$, and $\log((\alpha + \beta)z/\alpha)$ with terms up to order 2, 2, and 3 in the respective coordinates. The rational approximation was minimax fit to 2842 points in the remaining region for $0.01 < \alpha, \beta < 1000$. Test points were randomly sampled using log uniform sampling of α, β and stratified sampling of z conditioned on α, β. Minimax fitting achieved about half the maximum error of simple least squares fitting. Our complete approximation for $\frac{dz}{d\alpha}$ and $\frac{dz}{d\beta}$ is unit tested to have relative accuracy of 0.001 on a wide range of inputs.

8. Dirichlet Distribution

For completeness we record the general version of the formula for the pathwise gradient (given implicitly in the main text):

$$\frac{dz_i}{d\alpha_j} = -\frac{\frac{\partial F_{\rm Beta}}{\partial \alpha_j}(z_j|\alpha_j, \alpha_{\rm tot} - \alpha_j)}{{\rm Beta}(z_j|\alpha_j, \alpha_{\rm tot} - \alpha_j)} \times \left(\frac{\delta_{ij} - z_i}{1 - z_j}\right) \qquad (46)$$

We want to confirm that Eqn. 46 satisfies the transport equation for each choice of $j = 1, ..., n$:

$$\frac{\partial}{\partial \alpha_j} \log q + \nabla \cdot \boldsymbol{v} + \boldsymbol{v} \cdot \nabla \log q = 0 \qquad (47)$$

Treating z_j as a function of $\mathbf{z}_{-j} = (z_1, ..., z_{j-1}, z_{j+1}, ..., z_n)$ everywhere and introducing obvious shorthand for $F_{\rm Beta}(\cdot)$ and ${\rm Beta}(\cdot)$ we have:

$$\begin{aligned}
\nabla \cdot \boldsymbol{v} &= \sum_{i \neq j} \frac{\partial}{\partial z_i} \left(\frac{\frac{\partial F_{\rm Beta}}{\partial \alpha_j}(z_j|\alpha_j, \alpha_{\rm tot} - \alpha_j)}{{\rm Beta}(z_j|\alpha_j, \alpha_{\rm tot} - \alpha_j)} \frac{z_i}{\sum_{k \neq j} z_k} \right)\\
&= \frac{\frac{\partial F}{\partial \alpha_j}}{B} \frac{n - 2}{1 - z_j} - \frac{\partial \log B}{\partial \alpha_j} + \frac{\partial F}{\partial \alpha_j} \frac{(\log B)'}{B}
\end{aligned}$$

where $(\log B)'$ is differentiated w.r.t. the argument of $B(z_j)$. We further have that

$$\boldsymbol{v} \cdot \nabla \log q = \frac{\frac{\partial F}{\partial \alpha_j}}{B} \left(\sum_{i \neq j} \frac{\alpha_i - 1}{1 - z_j} - \frac{\alpha_j - 1}{z_j} \right)$$

and

$$\frac{\partial}{\partial \alpha_j} \log q = \psi(\alpha_j) - \psi(\alpha_{\rm tot}) + \log z_j$$

Since we have

$$\frac{\partial \log B}{\partial \alpha_j} = \psi(\alpha_j) - \psi(\alpha_{\rm tot}) + \log z_j$$

and

$$(\log B)' = \frac{\alpha_j - 1}{z_j} - \frac{\alpha_{\rm tot} - \alpha_j - 1}{1 - z_j}$$

it becomes clear by comparing the individual terms that everything cancels identically and so Eqn. 47 is in fact satisfied by the velocity field in Eqn. 46.

Finally, we note that Eqn. 46 is *not* the OMT solution in the coordinates $\boldsymbol{z}_{-j}$. It *is* the OMT solution in some coordinate system, but it is not readily apparent which coordinate system that might be.

9. Student's t-Distribution

As another example of how to compute pathwise gradients consider Student's t-distribution. Although we have not done so ourselves, it should be straightforward to compute an accurate approximation to Eqn. 1. In the absence of such an approximation, however, we can still get a pathwise gradient for the Student's t-distribution by composing the Normal and Gamma distributions:

$$\begin{aligned}
&\tau \sim {\rm Gamma}(\nu/2, 1) \qquad {\rm x}|\tau \sim \mathcal{N}(0, \tau^{-\frac{1}{2}})\\
&\Rightarrow z \equiv \sqrt{\tfrac{\nu}{2}} x \sim {\rm Student}(\nu)
\end{aligned} \qquad (48)$$

Since sampling z like this introduces an auxiliary random degree of freedom, pathwise gradients $\frac{dz}{d\nu}$ computed using Eqn. 48 will exhibit a larger variance than a direct computation of Eqn. 1 would yield.[8] The point is that *no additional work* is needed to obtain this particular form of the pathwise gradient: just use pathwise gradients for the Gamma and Normal distributions and the sampling procedure in Eqn. 48.

10. Baseball Experiment

To gain more insight into when we expect the OMT gradient estimator for the multivariate Normal distribution to outperform the RT gradient estimator, we conduct an additional experiment. We consider a model for repeated binary trial data (baseball players at bat) using the data in (Efron & Morris, 1975) and the modeling setup in (Stan Manual, 2017) with partial pooling. There are 18 baseball players and the data consists of 45 hits/misses for each player. The model has two global latent variables and 18 local latent variables so that the posterior is 20-dimensional. Specifically, the two global latent random variables are ϕ and κ, with priors $\mathrm{Uniform}(0,1)$ and $\mathrm{Pareto}(1, 1.5) \propto \kappa^{-5/2}$, respectively. The local latent random variables are given by θ_i for $i = 0, ..., 17$, with $p(\theta_i) = \mathrm{Beta}(\theta_\mathrm{i}|\alpha = \phi\kappa, \beta = (1-\phi)\kappa)$. The data likelihood factorizes into 45 Bernoulli observations with mean chance of success θ_i for each player i. The variational approximation is formed in the unconstrained space $\{\mathrm{logit}(\phi), \log(\kappa-1), \mathrm{logit}(\theta_\mathrm{i})\}$ and consists of a multivariate Normal distribution with a full-rank Cholesky factor $\boldsymbol{L}$. We use the Adam optimizer for training with a learning rate of 5×10^{-3} (Kingma & Ba, 2014).

For this particular model mean field SGVI performs reasonably well, since correlations between the latent random variables are not particularly strong. If we initialize $\boldsymbol{L}$ near the identity, we find that the OMT and RT gradient estimators perform nearly identically, with the difference that the former has an increased computational cost of about 25% per iteration. If, however, we initialize $\boldsymbol{L}$ far from the identity—so that the optimizer has to traverse a considerable distance in $\boldsymbol{L}$ space where the covariance matrix exhibits strong correlations—we find that the OMT estimator makes progress more quickly than the RT estimator and converges to a higher ELBO, see Fig. 5. Generalizing from this, we expect the OMT gradient estimator for the multivariate Normal distribution to exhibit better sample efficiency than the RT estimator in problems where the covariance matrix exhibits strong correlations. This is indeed the case for the GP experiment in the main text, where the learned kernel induces strong temporal correlations.

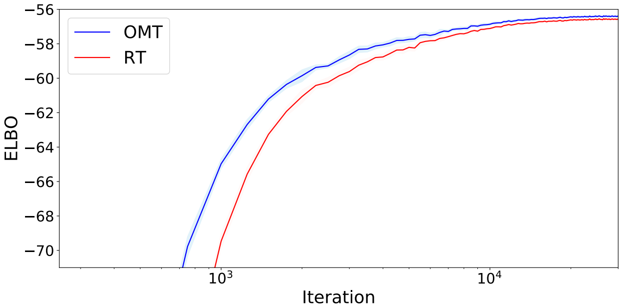

Figure 5. ELBO training curves for the experiment in Sec. 10 for the case where the Cholesky factor is initialized far from the identity. Depicted is the mean ELBO for 10 runs with $1-\sigma$ uncertainty bands around the mean. The OMT gradient estimator learns more quickly than the RT estimator and attains a higher ELBO.

11. Experimental Details

As noted in the main text, we use single-sample gradient estimators in all experiments. Unless noted otherwise, we always include the score function term for RSVI.

11.1. Multivariate Normal Synthetic Test Function Experiment

We describe the setup for the experiment corresponding to Fig. 5 in the main text. The dimension is fixed to $D = 50$ and the mean of $q_{\boldsymbol{\theta}}$ is fixed to the zero vector. The Cholesky factor $\boldsymbol{L}$ that enters into $q_{\boldsymbol{\theta}}$ is constructed as follows. The diagonal of $\boldsymbol{L}$ consists of all ones. To construct the off-diagonal terms we proceed as follows. We populate the entries below the diagonal of a matrix $\Delta\boldsymbol{L}$ by drawing each entry from the uniform distribution on the unit interval. Then we define $\boldsymbol{L} = \mathbb{1}_D + r\Delta\boldsymbol{L}$. Here r controls the magnitude of off-diagonal terms of $\boldsymbol{L}$ and appears on the horizontal axis of Fig. 5 in the main text. The three test functions are constructed as follows. First we construct a strictly lower diagonal matrix $\boldsymbol{Q}'$ by drawing each entry from a bernoulli distribution with probability 0.5. We then define $\boldsymbol{Q} = \boldsymbol{Q}' + \boldsymbol{Q}'^T$. The cosine test function is then given by

$$f(\boldsymbol{z}) = \cos\left(\sum_{i,j} Q_{ij} z_i / D\right) \tag{49}$$

The quadratic test function is given by

$$f(\boldsymbol{z}) = \boldsymbol{z}^T \boldsymbol{Q} \boldsymbol{z} \tag{50}$$

The quartic test function is given by

$$f(\boldsymbol{z}) = \left(\boldsymbol{z}^T \boldsymbol{Q} \boldsymbol{z}\right)^2 \tag{51}$$

In all cases the gradients can be computed analytically, which makes it easier to reliably estimate the variance of

[8] Note, however, that this additional variance will decrease as ν increases.

the gradient estimators.

11.2. Sparse Gamma DEF

Following (Naesseth et al., 2017), we use analytic expressions for each entropy term (as opposed to using the sampling estimate). We use the adaptive step sequence ρ^n proposed by (Kucukelbir et al., 2016) and also used in (Naesseth et al., 2017), which combines RMSPROP (Tieleman & Hinton, 2012) and Adagrad (Duchi et al., 2011):

$$\begin{aligned} \rho^n &= \eta \cdot n^{-1/2+\delta} \cdot \left(1 + \sqrt{s^n}\right)^{-1}. \\ s^n &= t\left(\hat{g}^n\right)^2 + (1-t)s^{n-1} \end{aligned} \tag{52}$$

Here $n = 1, 2, ...$ is the iteration number and the operations in Eqn. 52 are to be understood element-wise. In our case the gradient $\hat{g}^n$ is always a single-sample estimate. We fix $\delta = 10^{-16}$ and $t = 0.1$. In contrast to (Kucukelbir et al., 2016) but in line with (Naesseth et al., 2017) we initialize s_0 at zero. To choose η we did a grid search for each gradient estimator and each of the two model variants. Specifically, for each η we did 100 training iterations for three trials with different random seeds and then chose the η that yielded the highest mean ELBO after 100 iterations. This procedure led to the selection of $\eta = 4.5$ for the first model variant and $\eta = 30$ for the second model variant (note that within each model variant the gradient estimators preferred the same value of η). For the first model variant we included the score function-like term in the RSVI gradient estimator, while we did not include it for the second model variant, as we found that this hurt performance. In both cases we used the shape augmentation setting $B = 4$, which was also used for the results reported in (Naesseth et al., 2017). After fixing η we trained the model for 2000 iterations, initializing with another random number seed. The figure in the main text shows the training curves for that single run. We confirmed that other random number seeds give similar results. A reference implementation can be found here: https://github.com/uber/pyro/blob/0.2.1/examples/sparse_gamma_def.py

11.3. Gaussian Process Regression

We used the Adam optimizer (Kingma & Ba, 2014) to optimize the ELBO with single-sample gradient estimates. We chose the Adam hyperparameters by doing a grid search over the learning rate and β_1. For each combination (lr, β_1) we did 20 training iterations for three trials with different random seeds and then chose the combination that yielded the highest mean ELBO after 20 iterations. This procedure led to the selection of a learning rate of 0.030 and $\beta_1 = 0.50$ for both gradient estimators (OMT and reparameterization trick). We then trained the model for 500 iterations, initializing with another random number seed. The figure in the main text shows the training curves for that single run. We confirmed that other random number seeds give similar results.

References

Butler, Ronald W. *Saddlepoint approximations with applications*, volume 22. Cambridge University Press, 2007.

Duchi, John, Hazan, Elad, and Singer, Yoram. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12 (Jul):2121–2159, 2011.

Efron, Bradley and Morris, Carl. Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

Kingma, Diederik P and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kucukelbir, Alp, Tran, Dustin, Ranganath, Rajesh, Gelman, Andrew, and Blei, David M. Automatic differentiation variational inference. *arXiv preprint arXiv:1603.00788*, 2016.

Lugannani, Robert and Rice, Stephen. Saddle point approximation for the distribution of the sum of independent random variables. *Advances in applied probability*, 12 (2):475–490, 1980.

Naesseth, Christian, Ruiz, Francisco, Linderman, Scott, and Blei, David. Reparameterization gradients through acceptance-rejection sampling algorithms. In *Artificial Intelligence and Statistics*, pp. 489–498, 2017.

Stan Manual. Stan modeling language users guide and reference manual, version 2.17.0. http://mc-stan.org/users/documentation/case-studies/pool-binary-trials.html, 2017.

Tieleman, T. and Hinton, G. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.