Pathwise Derivatives Beyond the Reparameterization Trick

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

Abstract

We observe that gradients computed via the reparameterization trick are in direct correspondence with solutions of the transport equation in the formalism of optimal transport. We use this perspective to compute (approximate) pathwise gradients for probability distributions not directly amenable to the reparameterization trick: Gamma, Beta, and Dirichlet. We further observe that when the reparameterization trick is applied to the Cholesky-factorized multivariate Normal distribution, the resulting gradients are suboptimal in the sense of optimal transport. We derive the optimal gradients and show that they have reduced variance in a Gaussian Process regression task. We demonstrate with a variety of synthetic experiments and stochastic variational inference tasks that our pathwise gradients are competitive with other methods.

1. Introduction

Maximizing objective functions via gradient methods is ubiquitous in machine learning. When the objective function $\mathcal{L}$ is defined as an expectation of a (differentiable) test function $f_{\boldsymbol{\theta}}(\boldsymbol{z})$ w.r.t. a probability distribution $q_{\boldsymbol{\theta}}(\boldsymbol{z})$,

$$\mathcal{L} = \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})} \left[f_{\boldsymbol{\theta}}(\boldsymbol{z}) \right] \tag{1}$$

computing exact gradients w.r.t. the parameters $\boldsymbol{\theta}$ is often unfeasible so that optimization methods must instead make due with stochastic gradient estimates. If the gradient estimator is unbiased, then stochastic gradient descent with an appropriately chosen sequence of step sizes can be shown to have nice convergence properties (Robbins & Monro, 1951). If, however, the gradient estimator exhibits large variance, stochastic optimization algorithms may be impractically slow. Thus it is of general interest to develop gradient estimators with reduced variance.

We revisit the class of gradient estimators popularized in (Kingma & Welling, 2013; Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014), which go under the name of the pathwise derivative or the reparameterization trick. While this class of gradient estimators is not applicable to all choices of probability distribution $q_{\boldsymbol{\theta}}(\boldsymbol{z})$, empirically it has been shown to yield suitably low variance in many cases of practical interest and thus has seen wide use. We show that the pathwise derivative in the literature is in fact a particular instance of a continuous family of gradient estimators. Drawing a connection to tangent fields in the field of optimal transport,[1] we show that one can define a unique pathwise gradient that is optimal in the sense of optimal transport. For the purposes of this paper, we will refer to these optimal gradients as OMT (optimal mass transport) gradients.

The resulting geometric picture is particularly intriguing in the case of multivariate distributions, where each choice of gradient estimator specifies a velocity field on the sample space. To make this picture more concrete, in Figure 1 we show the velocity fields that correspond to two different gradient estimators for the off-diagonal element of the Cholesky factor parameterizing a bivariate Normal distribution. We note that the velocity field that corresponds to the reparameterization trick has a large rotational component that makes it suboptimal in the sense of optimal transport. In Sec. 7 we show that this suboptimality can result in reduced performance when fitting a Gaussian Process to data.

The rest of this paper is organized as follows. In Sec. 2 we provide a brief overview of stochastic gradient variational inference. In Sec. 3 we show how to compute pathwise gradients for univariate distributions. In Sec. 4 we expand our discussion of pathwise gradients to the case of multivariate distributions, introduce the connection to the transport equation, and provide an analytic formula for the OMT gradient in the case of the multivariate Normal. In Sec. 5 we discuss how we can compute high precision approximate pathwise gradients for the Gamma, Beta, and Dirichlet distributions. In Sec. 6 we place our work in the context of related research. In Sec. 7 we demonstrate the performance of our gradient estimators with a variety of synthetic experiments and experiments on real world datasets. Finally, in Sec. 8 we conclude with a discussion of directions for future work.

[*]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.

[1]See (Villani, 2003; Ambrosio et al., 2008) for a review.

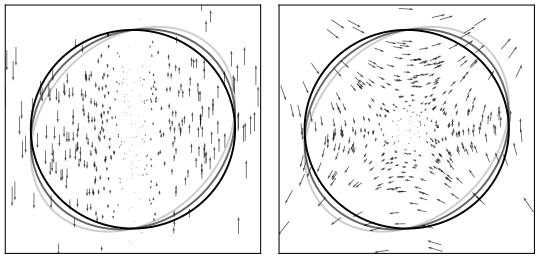

Figure 1. Velocity fields for a bivariate Normal distribution parameterized by a Cholesky factor $\boldsymbol{L} = \mathbb{1}_2$. The gradient is w.r.t. the off-diagonal element L_{21}. On the left we depict the velocity field corresponding to the reparameterization trick and on the right we depict the velocity field that is optimal in the sense of optimal transport. The solid black circle denotes the 1-σ covariance ellipse, with the gray ellipses denoting displaced covariance ellipses that result from small increases in L_{21}. Note that the ellipses evolve the same way under both velocity fields, but *individual* particles flow differently to effect the same global displacement of mass.

2. Stochastic Gradient Variational Inference

One area where stochastic gradient estimators play a particularly central role is stochastic variational inference (Hoffman et al., 2013). This is especially the case for black-box methods (Wingate & Weber, 2013; Ranganath et al., 2014), where conjugacy and other simplifying structural assumptions are unavailable, with the consequence that Monte Carlo estimators become necessary. For concreteness, we will refer to this class of methods as Stochastic Gradient Variational Inference (SGVI). In this section we give a brief overview of this line of research, as it serves as the motivating use case for our work. Furthermore, in Sec. 7 SGVI will serve as the main testbed for our proposed methods.

Let $p(\boldsymbol{x}, \boldsymbol{z})$ define a joint probability distribution over observed data $\boldsymbol{x}$ and latent random variables $\boldsymbol{z}$. One of the main tasks in Bayesian inference is to compute the posterior distribution $p(\boldsymbol{z}|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{x})}$. For many models of interest, this is an intractably hard problem and so approximate methods become necessary. Variational inference recasts Bayesian inference as an optimization problem. Specifically we define a variational family of distributions $q_{\boldsymbol{\theta}}(\boldsymbol{z})$ parameterized by $\boldsymbol{\theta}$ and seek to find a value of $\boldsymbol{\theta}$ that minimizes the KL divergence between $q_{\boldsymbol{\theta}}(\boldsymbol{z})$ and the (unknown) posterior $p(\boldsymbol{z}|\boldsymbol{x})$. This is equivalent to maximizing the ELBO (Jordan et al., 1999), defined as

$$\text{ELBO} = \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})}\left[\log p(\boldsymbol{x}, \boldsymbol{z}) - \log q_{\boldsymbol{\theta}}(\boldsymbol{z})\right] \tag{2}$$

For general choices of $p(\boldsymbol{x}, \boldsymbol{z})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{z})$, this expectation—much less its gradients—cannot be computed analytically. In these circumstances a natural approach is to build a Monte Carlo estimator of the ELBO and its gradient w.r.t. $\boldsymbol{\theta}$. The properties of the chosen gradient estimator—especially its bias and variance—play a critical rule in determining the viability of the resulting stochastic optimization. Next, we review two commonly used gradient estimators; we leave a brief discussion of more elaborate variants to Sec. 6.

2.1. Score Function Estimator

The score function estimator, also referred to as the log-derivative trick or REINFORCE (Glynn, 1990; Williams, 1992), provides a simple and broadly applicable recipe for estimating ELBO gradients (Paisley et al., 2012). The score function estimator expresses the gradient as an expectation with respect to $q_{\boldsymbol{\theta}}(\boldsymbol{z})$, with the simplest variant given by

$$\nabla_{\boldsymbol{\theta}}\text{ELBO} = \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})}\left[\nabla_{\boldsymbol{\theta}} \log r + \log r \nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{z})\right] \tag{3}$$

where $\log r = \log p(\boldsymbol{x}, \boldsymbol{z}) - \log q_{\boldsymbol{\theta}}(\boldsymbol{z})$. Monte Carlo estimates of Eqn. 3 can be formed by drawing samples from $q_{\boldsymbol{\theta}}(\boldsymbol{z})$ and computing the term in the square brackets. Although the score function estimator is very general (e.g. it applies to discrete random variables) it typically suffers from high variance, although this can be mitigated with the use of variance reduction techniques such as Rao-Blackwellization (Casella & Robert, 1996) and control variates (Ross, 2006).

2.2. Pathwise Gradient Estimator

The pathwise gradient estimator, a.k.a. the reparameterization trick (RT), is not as broadly applicable as the score function estimator, but it generally exhibits lower variance (Price, 1958; Salimans et al., 2013; Kingma & Welling, 2013; Glasserman, 2013; Rezende et al., 2014; Titsias & Lázaro-Gredilla, 2014). It is applicable to continuous random variables whose probability density $q_{\boldsymbol{\theta}}(\boldsymbol{z})$ can be reparameterized such that we can rewrite expectations

$$\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})}\left[f_{\boldsymbol{\theta}}(\boldsymbol{z})\right] \quad \longrightarrow \quad \mathbb{E}_{q_0(\boldsymbol{\epsilon})}\left[f_{\boldsymbol{\theta}}(\mathcal{T}(\boldsymbol{\epsilon}; \boldsymbol{\theta}))\right] \tag{4}$$

where $q_0(\boldsymbol{z})$ is a fixed distribution with no dependence on $\boldsymbol{\theta}$ and $\mathcal{T}(\boldsymbol{\epsilon}; \boldsymbol{\theta})$ is a differentiable $\boldsymbol{\theta}$-dependent transformation. Since the expectation w.r.t. $q_0(\boldsymbol{\epsilon})$ has no $\boldsymbol{\theta}$ dependence, gradients w.r.t. $\boldsymbol{\theta}$ can be computed by pushing $\nabla_{\boldsymbol{\theta}}$ through the expectation. This reparameterization can be done for a number of distributions, including for example the Normal distribution. Unfortunately the reparameterization trick is non-trivial to apply to a number of commonly used distributions, e.g. the Gamma and Beta distributions, since the required shape transformations $\mathcal{T}(\boldsymbol{\epsilon}; \boldsymbol{\theta})$ inevitably involve special functions.

3. Univariate Pathwise Gradients

Consider an objective function given as the expectation of a test function $f_{\boldsymbol{\theta}}(z)$ with respect to a distribution $q_{\boldsymbol{\theta}}(z)$,

where z is a continuous one-dimensional random variable:

$$\mathcal{L} = \mathbb{E}_{q_{\boldsymbol{\theta}}(z)}\left[f_{\boldsymbol{\theta}}(z)\right] \tag{5}$$

Here $q_{\boldsymbol{\theta}}(z)$ and $f_{\boldsymbol{\theta}}(z)$ are parameterized by $\boldsymbol{\theta}$, and we would like to compute (stochastic) gradients of $\mathcal{L}$ w.r.t. θ, where θ is a scalar component of $\boldsymbol{\theta}$:

$$\nabla_\theta \mathcal{L} = \nabla_\theta \mathbb{E}_{q_{\boldsymbol{\theta}}(z)}\left[f_{\boldsymbol{\theta}}(z)\right] \tag{6}$$

Crucially we would like to avoid the log-derivative trick, which yields a gradient estimator that tends to have high variance. Doing so will be easy if we can rewrite the expectation in terms of a fixed distribution that does not depend on θ. A natural choice is to use the standard uniform distribution $\mathcal{U}$,

$$\mathcal{L} = \mathbb{E}_{\mathcal{U}(u)}\left[f_{\boldsymbol{\theta}}(F^{-1}_{\boldsymbol{\theta}}(u))\right] \tag{7}$$

where the transformation $F^{-1}_{\boldsymbol{\theta}} : u \to z$ is the inverse CDF of $q_{\boldsymbol{\theta}}(z)$. As desired, all dependence on θ is now inside the expectation. Unfortunately, for many continuous univariate distributions of interest (e.g. the Gamma and Beta distributions) the transformation $F^{-1}_{\boldsymbol{\theta}}$ (as well as its derivative w.r.t. $\boldsymbol{\theta}$) does not admit a simple analytic expression.

Fortunately, by making use of implicit differentiation we can compute the gradient in Eqn. 6 without explicitly introducing $F^{-1}_{\boldsymbol{\theta}}$. To complete the derivation define u by

$$u \equiv F_{\boldsymbol{\theta}}(z) = \int_{-\infty}^{z} q_{\boldsymbol{\theta}}(z')dz' \tag{8}$$

and differentiate both sides of Eqn. 8 w.r.t. θ and make use of the fact that $u \sim \mathcal{U}$ does not depend on θ to obtain

$$0 = \frac{dz}{d\theta} q_{\boldsymbol{\theta}}(z) + \int_{-\infty}^{z} \frac{\partial}{\partial \theta} q_{\boldsymbol{\theta}}(z')dz' \tag{9}$$

This then yields our master formula for the univariate case

$$\frac{dz}{d\theta} = -\frac{\frac{\partial F_{\boldsymbol{\theta}}}{\partial \theta}(z)}{q_{\boldsymbol{\theta}}(z)} \tag{10}$$

where the corresponding gradient estimator is given by

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{q_{\boldsymbol{\theta}}(z)}\left[\frac{df_{\boldsymbol{\theta}}(z)}{dz}\frac{dz}{d\theta} + \frac{\partial f_{\boldsymbol{\theta}}(z)}{\partial \theta}\right] \tag{11}$$

While this derivation is elementary, it helps to clarify things: the key ingredient needed to compute pathwise gradients in Eqn. 6 is the ability to compute (or approximate) the derivative of the CDF, i.e. $\frac{\partial}{\partial \theta} F_{\boldsymbol{\theta}}(z)$. In the supplementary materials we verify that Eqn. 11 results in correct gradients.

It is worth emphasizing how this approach differs from a closely related alternative. Suppose we construct a (differentiable) approximation of the *inverse* CDF, $\hat{F}^{-1}_{\boldsymbol{\theta}}(u) \approx F^{-1}_{\boldsymbol{\theta}}(u)$. For example, we might train a neural network $\mathrm{nn}(u, \boldsymbol{\theta}) \approx F^{-1}_{\boldsymbol{\theta}}(u)$. We can then push samples $u \sim \mathcal{U}$ through $\mathrm{nn}(u, \boldsymbol{\theta})$ and obtain approximate samples from $q_{\boldsymbol{\theta}}(z)$ as well as approximate derivatives $\frac{dz}{d\theta}$ via the chain rule; in this case, there will be a mismatch between the probability $q_{\boldsymbol{\theta}}(z)$ assigned to samples z and the actual distribution over z. By contrast, if we use the construction of Eqn. 10, our samples z will still be exact[2] and the fidelity of our approximation of (the derivatives of) $F_{\boldsymbol{\theta}}(z)$ will only affect the accuracy of our approximation for $\frac{dz}{d\theta}$.

4. Multivariate Pathwise Gradients

In the previous section we focused on continuous univariate distributions. Pathwise gradients can also be constructed for continuous multivariate distributions, although the analysis is in general expected to be much more complicated than in the univariate case—directly analogous to the difference between ordinary and partial differential equations. Before constructing estimators for particular distributions, we introduce the connection to the transport equation.

4.1. The Transport Equation

Consider a multivariate distribution $q_{\boldsymbol{\theta}}(\boldsymbol{z})$ in D dimensions and consider differentiating $\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})}\left[f(\boldsymbol{z})\right]$ with respect to the parameter θ.[3] As we vary θ we move $q_{\boldsymbol{\theta}}(\boldsymbol{z})$ along a curve in the space of distributions over the sample space. Alternatively, we can think of each distribution as a cloud of particles; as we vary θ from θ to $\theta + \Delta\theta$ each particle undergoes an infinitesimal displacement $d\boldsymbol{z}$. Any set of displacements that ensures that the displaced particles are distributed according to the displaced distribution $q_{\theta+\Delta\theta}(\boldsymbol{z})$ is allowed. This intuitive picture can be formalized with the transport a.k.a. continuity equation:[4]

$$\frac{\partial}{\partial \theta} q_{\boldsymbol{\theta}} + \nabla_{\boldsymbol{z}} \cdot \left(q_{\boldsymbol{\theta}} \boldsymbol{v}^\theta\right) = 0 \tag{12}$$

Here the *velocity field* $\boldsymbol{v}^\theta$ is a vector field defined on the sample space that displaces samples (i.e. particles) $\boldsymbol{z}$ as we vary θ infinitesimally. Note that there is a velocity field $\boldsymbol{v}^\theta$ for each component θ of $\boldsymbol{\theta}$. This equation is readily interpreted in the language of fluid dynamics. In order for the the total probability to be conserved, the term $\frac{\partial}{\partial \theta} q_{\boldsymbol{\theta}}(\boldsymbol{z})$—which is the rate of change of the number of particles in the infinitesimal volume element at $\boldsymbol{z}$—has to be counterbalanced by the in/out-flow of particles—as given by the divergence term.

[2] Or rather their exactness will be determined by the quality of our sampler for $q_{\boldsymbol{\theta}}(z)$, which is fully decoupled from how we compute derivatives $\frac{dz}{d\theta}$.

[3] Here without loss of generality we assume that $f(\boldsymbol{z})$ has no dependence on θ, since computing $\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})}\left[\nabla_\theta f_{\boldsymbol{\theta}}(\boldsymbol{z})\right]$ presents no difficulty; the difficulty stems from the dependence on θ in $q_{\boldsymbol{\theta}}(z)$.

[4] We refer the reader to (Villani, 2003) and (Ambrosio et al., 2008) for details.

4.2. Gradient Estimator

Given a solution to Eqn. 12, we can form the gradient estimator

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})}\left[\boldsymbol{v}^\theta \cdot \nabla_{\mathbf{z}} f\right] \tag{13}$$

which generalizes Eqn. 11 to the multivariate case. That this is an unbiased gradient estimator follows directly from the divergence theorem (see the supplementary materials).

4.3. Tangent Fields

In general Eqn. 12 admits an infinite dimensional space of solutions. In the context of our derivation of Eqn. 10, we might loosely say that different solutions of Eqn. 12 correspond to different ways of specifying quantiles of $q_{\boldsymbol{\theta}}(\boldsymbol{z})$. To determine a *unique*[5] solution—the tangent field from the theory of optimal transport—we require that

$$\frac{\partial v_i^{\mathrm{OMT}}}{\partial z_j} = \frac{\partial v_j^{\mathrm{OMT}}}{\partial z_i} \qquad \forall i, j \tag{14}$$

In this case it can be shown that $\boldsymbol{v}^{\mathrm{OMT}}$ minimizes the total kinetic energy, which is given by[6]

$$K(\boldsymbol{v}) = \tfrac{1}{2}\int d\boldsymbol{z}\; q_{\boldsymbol{\theta}}(z)||\boldsymbol{v}||^2 \tag{15}$$

4.4. Gradient variance

The $||\boldsymbol{v}||^2$ term that appears in Eqn. 15 might lead one to hope that $\boldsymbol{v}^{\mathrm{OMT}}$ provides gradients that minimize gradient variance. Unfortunately, the situation is more complicated. Denoting the (mean) gradient by $\boldsymbol{g} = \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})}[\boldsymbol{v} \cdot \nabla_{\mathbf{z}} f(\boldsymbol{z})]$ the total gradient variance is given by

$$\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{z})}\left[||\boldsymbol{v} \cdot \nabla_{\mathbf{z}} f||^2\right] - ||\boldsymbol{g}||^2 \tag{16}$$

Since $\boldsymbol{g}$ is the same for all unbiased gradient estimators, the gradient estimator that minimizes the total variance is the one that minimizes the first term in Eqn. 16. For test functions $f(\boldsymbol{z})$ that approximately satisfy $\nabla_{\mathbf{z}} f \propto \mathbb{1}$ over the bulk of the support of $q_{\boldsymbol{\theta}}(\boldsymbol{z})$, the first term in Eqn. 16 term is approximately proportional to the kinetic energy. In this case the OMT gradient estimator will be (nearly) optimal. Note that the kinetic energy weighs contributions from different components of $\boldsymbol{v}$ equally, whereas $\boldsymbol{g}$ scales different components of $\boldsymbol{v}$ with $\nabla_{\mathbf{z}} f$. Thus we can think of the OMT gradient estimator as a good choice for generic choices of $f(\boldsymbol{z})$ that are relatively flat and isotropic (or, alternatively, for choices of $f(\boldsymbol{z})$ where we have little *a priori* knowledge about the detailed structure of $\nabla_{\mathbf{z}} f$). So for any particular choice of a generic $f(\boldsymbol{z})$ there will be some gradient estimator that has lower variance than the OMT gradient estimator. Still, for *many* choices of $f(\boldsymbol{z})$ we expect the OMT gradient estimator to have lower variance than the RT gradient estimator, since the latter has no particular optimality guarantees (at least not in any coordinate system that we expect to be well adapted to $f(\boldsymbol{z})$).

4.5. The Multivariate Normal

In the case of a (zero mean) multivariate Normal distribution parameterized by a Cholesky factor $\boldsymbol{L}$ via $\boldsymbol{z} = \boldsymbol{L}\tilde{\boldsymbol{z}}$, where $\tilde{\boldsymbol{z}}$ is white noise, the reparameterization trick yields the following velocity field for L_{ab}:[7]

$$v_i^{\mathrm{RT}} = \frac{\partial z_i}{\partial L_{ab}} = \delta_{ia}(L^{-1}\boldsymbol{z})_b \tag{17}$$

Note that Eqn. 17 is just a particular instance of the solution to the transport equation that is implicitly provided by the reparameterization trick, namely

$$\boldsymbol{v}^\theta = \left.\frac{\partial \mathcal{T}(\boldsymbol{\epsilon};\boldsymbol{\theta})}{\partial \theta}\right|_{\boldsymbol{\epsilon}=\mathcal{T}^{-1}(\boldsymbol{z};\boldsymbol{\theta})} \tag{18}$$

In the supplementary materials we verify that Eqn. 17 satisfies the transport equation Eqn. 12. However, it is evidently *not* optimal in the sense of optimal transport, since $\frac{\partial v_i^{\mathrm{RT}}}{\partial z_j} = \delta_{ia} L_{bj}^{-1}$ is not symmetric in i and j. In fact the tangent field takes the form

$$v_i^{\mathrm{OMT}} = \frac{1}{2}\left(\delta_{ia}(L^{-1}\boldsymbol{z})_b + z_a L_{bi}^{-1}\right) + (S^{ab}\boldsymbol{z})_i \tag{19}$$

where S^{ab} is a symmetric matrix whose precise form we give in the supplementary materials. We note that computing gradients with Eqn. 19 is $\mathcal{O}(D^3)$, since it involves a singular value decomposition of the covariance matrix. In Sec. 7 we show that the resulting gradient estimator can lead to reduced variance.

5. Numerical Recipes

In this section we show how Eqn. 10 can be used to obtain pathwise gradients in practice. In many cases of interest we will need to derive approximations to $\frac{\partial}{\partial\theta}F(z)$ that balance the need for high accuracy (thus yielding gradient estimates with negligible bias) with the need for computational efficiency. In particular we will derive accurate approximations to Eqn. 10 for the Gamma, Beta, and Dirichlet distributions. These approximations will involve three basic components:

1. Elementary Taylor expansions
2. The Lugannani-Rice saddlepoint expansion (Lugannani & Rice, 1980; Butler, 2007)

[5] We refer the reader to Ch. 8 of (Ambrosio et al., 2008) for details.

[6] Note that the univariate solution, Eqn. 10, is automatically the OMT solution.

[7] Note that the reparameterization trick already yields the OMT gradient for the location parameter $\boldsymbol{\mu}$.

3. Rational polynomial approximations in regions of (z, θ) that are analytically intractable

5.1. Gamma

The CDF of the Gamma distribution involves the (lower) incomplete gamma function $\gamma(\cdot)$: $F_{\alpha,\beta}(z) = \frac{\gamma(\alpha,\beta z)}{\Gamma(\alpha)}$. Unfortunately $\gamma(\cdot)$ does not admit simple analytic expressions for derivatives w.r.t. its first argument, and so we must resort to numerical approximations. Since $z \sim \text{Gamma}(\alpha, \beta = 1) \Leftrightarrow \text{z}/\beta \sim \text{Gamma}(\alpha, \beta)$ it is sufficient to consider $\frac{dz}{d\alpha}$ for the standard Gamma distribution with $\beta = 1$.

5.1.1. $z \ll 1$

To give a flavor for the kinds of approximations we use, consider how we can approximate $\frac{\partial}{\partial\alpha}\gamma(\alpha, z)$ in the limit $z \ll 1$. We simply do a Taylor series in powers of z:

$$\begin{aligned}\frac{\partial}{\partial\alpha}\gamma(\alpha, z) &= \frac{\partial}{\partial\alpha}\int_0^z (z')^\alpha \left(1/z' - 1 + \tfrac{1}{2}z' + ...\right) dz' \\ &= \frac{\partial}{\partial\alpha} z^\alpha \left(\frac{1}{\alpha} - \frac{z}{\alpha+1} + \frac{\frac{1}{2}z^2}{\alpha+2} + ...\right)\end{aligned}$$

In practice we use 6 terms in this expansion, which is accurate for $z < 0.8$. Details for the remaining approximations can be found in the supplementary materials.

5.2. Beta

The CDF of the Beta distribution, F_{Beta}, is the (regularized) incomplete beta function; just like in the case of the Gamma distribution, its derivatives do not admit simple analytic expressions. We describe the numerical approximations we used in the supplementary materials.

5.3. Dirichlet

Let $\boldsymbol{z} \sim \text{Dir}(\boldsymbol{\alpha})$ be Dirichlet distributed with n components. Noting that the z_i are constrained to lie within the unit $(n-1)$-simplex, we proceed by representing $\boldsymbol{z}$ in terms of $n-1$ mutually independent Beta variates (Wilks, 1962):

$$\begin{aligned}&\tilde{z}_i \sim \text{Beta}(\alpha_\text{i}, \textstyle\sum_{\text{j=i+1}}^{\text{n}}\alpha_\text{j}) \qquad \text{for} \qquad \text{i} = 1, ..., \text{n}-1 \\ &z_1 = \tilde{z}_1 \qquad\qquad z_n = \textstyle\prod_{j=1}^{n-1}(1-\tilde{z}_j) \\ &z_i = \tilde{z}_i \textstyle\prod_{j=1}^{i-1}(1-\tilde{z}_j) \qquad \text{for} \qquad \text{i} = 2, ..., \text{n}-1\end{aligned}$$

Without loss of generality, we will compute $\frac{d}{d\alpha_1}z_i$ for $i = 1, ..., n$. Crucially, the only dependence on α_1 in Eqn. 20 is through $\tilde{z}_1$. We find:

$$\frac{d\boldsymbol{z}}{d\alpha_1} = -\frac{\frac{\partial F_{\text{Beta}}}{\partial\alpha_1}(z_1|\alpha_1, \alpha_\text{tot} - \alpha_1)}{\text{Beta}(\text{z}_1|\alpha_1, \alpha_\text{tot} - \alpha_1)} \times \left(1, \frac{-z_2}{1-z_1}, ..., \frac{-z_n}{1-z_1}\right) \tag{20}$$

Note that Eqn. 20 implies that $\frac{d}{d\boldsymbol{\alpha}}\sum_i z_i = 0$, as it must because of the simplex constraint. Since we have already developed an approximation for $\frac{\partial F_{\text{Beta}}}{\partial\theta}$, Eqn. 20 provides a complete recipe for pathwise Dirichlet gradients. Note that although we have used a stick-breaking construction to derive Eqn. 20, this in no way dictates the sampling scheme we use when generating $\boldsymbol{z} \sim \text{Dir}(\boldsymbol{\alpha})$. In the supplementary materials we verify that Eqn. 20 satisfies the transport equation.

5.4. Implementation

It is worth emphasizing that pathwise gradient estimators of the form in Eqn. 13 have the advantage of being 'plug-and-play.' We simply plug an approximate or exact velocity field into our favorite automatic differentiation engine[8] so that samples $\boldsymbol{z}$ and $f_{\boldsymbol{\theta}}(\boldsymbol{z})$ are differentiable w.r.t. $\boldsymbol{\theta}$. There is no need to construct a surrogate objective function to form the gradient estimator.

6. Related Work

A number of lines of research bears upon our work. There is a large body of work on constructing gradient estimators with reduced variance, much of which can be understood in terms of control variates (Ross, 2006): for example, (Mnih & Gregor, 2014) construct neural baselines for score-function gradients; (Schulman et al., 2015) discuss gradient estimators for stochastic computation graphs and their Rao-Blackwellization; and (Tucker et al., 2017; Grathwohl et al., 2017) construct adaptive control variates for discrete random variables. Another example of this line of work is reference (Miller et al., 2017), where the authors construct control variates that are applicable when $q_{\boldsymbol{\theta}}(\boldsymbol{z})$ is a *diagonal* Normal distribution. While our OMT gradient for the multivariate Normal distribution, Eqn. 19, can also be understood in the language of control variates,[9] (Miller et al., 2017) relies on Taylor expansions of the test function $f_{\boldsymbol{\theta}}(\boldsymbol{z})$.[10]

In (Graves, 2016), the author derives formula Eqn. 10 and uses it to construct gradient estimators for mixture distributions. Unfortunately, the resulting gradient estimator is expensive, relying on a recursive computation that scales with the dimension of the sample space.

Another line of work constructs partially reparameterized gradient estimators for cases where the reparameterization trick is difficult to apply. The generalized reparameterization gradient (G-REP) (Ruiz et al., 2016) uses standardization

[8] Our approximations for pathwise gradients for the Gamma, Beta, and Dirichlet distributions are available in the 0.4 release of PyTorch (Paszke et al., 2017).

[9] See Sec. 8 and the supplementary materials for a brief discussion.

[10] In addition, note that in their approach variance reduction for gradients w.r.t. the scale parameter $\boldsymbol{\sigma}$ necessitates a multi-sample estimator (at least for high-dimensional models where computing the diagonal of the Hessian is prohibitively expensive).

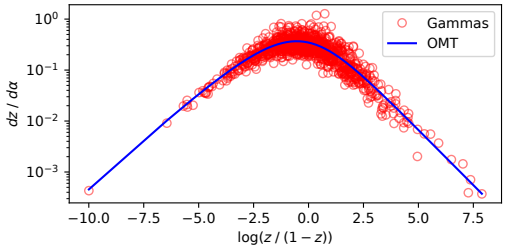

Figure 2. Derivatives $\frac{dz}{d\alpha}$ for samples $z \sim \text{Beta}(1, 1)$. We compare the OMT gradient to the gradient that is obtained when samples $z \sim \text{Beta}(\alpha, \beta)$ are represented as the ratio of two Gamma variates (each with its own pathwise derivative). The OMT derivative has a deterministic value for each sample z, whereas the Gamma representation induces a higher variance stochastic derivative due to the presence of an auxiliary random variable.

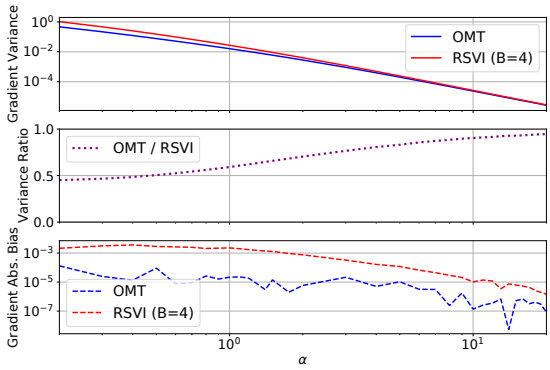

Figure 3. We compare the OMT gradient to the RSVI gradient with $B = 4$ for the test function $f(z) = z^3$ and $q_{\boldsymbol{\theta}}(z) = \text{Beta}(z|\alpha, \alpha)$. In the bottom panel we depict finite-sample bias for 25 million samples (this also includes effects from finite numerical precision).

via sufficient statistics to obtain a transformation $\mathcal{T}(\boldsymbol{\epsilon}; \boldsymbol{\theta})$ that minimizes the dependence of $q(\boldsymbol{\epsilon})$ on $\boldsymbol{\theta}$. This results in a partially reparameterized gradient estimator that also includes a score function-like term.[11] In RSVI (Naesseth et al., 2017) the authors consider gradient estimators in the case that $q_{\boldsymbol{\theta}}(\boldsymbol{z})$ can be sampled from efficiently via rejection sampling. This results in a gradient estimator with the same generic structure as G-REP, although in the case of RSVI the score function-like term can often be dropped in practice at the cost of small bias (with the benefit of reduced variance). Besides the fact that this gradient estimator is not fully pathwise, one key difference with our approach is that for many distributions of interest (e.g. the Beta and Dirichlet distributions), rejection sampling introduces auxiliary random variables, which results in additional stochasticity and thus higher variance (cf. Figure 2). In contrast our pathwise gradients for the Beta and Dirichlet distributions are *deterministic* for a given $\boldsymbol{z}$ and $\boldsymbol{\theta}$. Finally, (Knowles, 2015) uses (somewhat imprecise) approximations to the inverse CDF to derive gradient estimators for Gamma random variables.

As the final version of this manuscript was being prepared, we became aware of (Figurnov et al., 2018), which has some overlap with this work. In particular, (Figurnov et al., 2018) derives Eqn. 10 and an interesting generalization to the multivariate case. This allows the authors to construct pathwise derivatives for the Gamma, Beta, and Dirichlet distributions. For the latter two distributions, however, the derivatives include additional stochasticity that our pathwise derivatives avoid. Also, the authors do not draw the connection to the transport equation and optimal transport or consider the multivariate Normal distribution in any detail.

[11] That is a term in the gradient estimator that is proportional to the test function $f_{\boldsymbol{\theta}}(\boldsymbol{z})$.

7. Experiments

All experiments in this section use single-sample gradient estimators.

7.1. Synthetic Experiments

In this section we validate our pathwise gradients for the Beta, Dirichlet, and multivariate Normal distributions. Where appropriate we compare to the RT gradient, the score function gradient, or RSVI.

7.1.1. BETA DISTRIBUTION

In Fig. 3 we compare the performance of our OMT gradient for Beta random variables to the RSVI gradient estimator. We use a test function $f(z) = z^3$ for which we can compute the gradient exactly. We see that the OMT gradient performs favorably over the entire range of parameter α that defines the distribution $\text{Beta}(\alpha, \alpha)$ used to compute $\mathcal{L}$. For smaller α, where $\mathcal{L}$ exhibits larger curvature, the variance of the estimator is noticeably reduced. Notice that one reason for the reduced variance of the OMT estimator as compared to the RSVI estimator is the presence of an auxiliary random variable in the latter case (cf. Figure 2).

7.1.2. DIRICHLET DISTRIBUTION

In Fig. 4 we compare the variance of our pathwise gradient for the Dirichlet distribution to the RSVI gradient estimator. We compute stochastic gradients of the ELBO for a Multinomial-Dirichlet model initialized at the exact posterior (where the exact gradient is zero). The Dirichlet distribution has 1995 components, and the single data point is a bag of words from a natural language document. We see that the pathwise gradient performs favorably over the entire

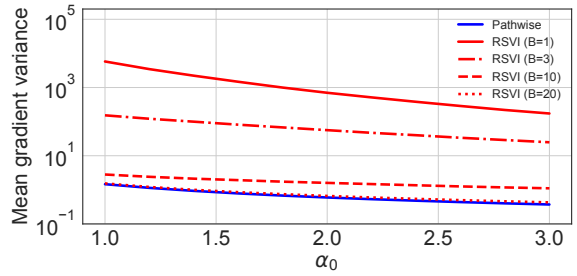

Figure 4. Gradient variance for the ELBO of a conjugate Multinomial-Dirichlet model. We compare the pathwise gradient to RSVI for different boosts B. See Sec. 7.1.2 for details.

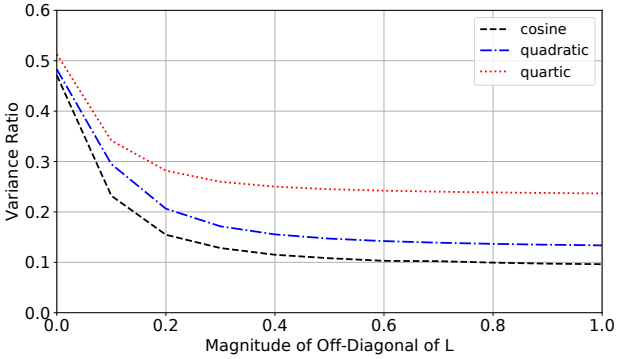

Figure 5. We compare the OMT gradient estimator for the multivariate Normal distribution to the RT estimator for three test functions. The horizontal axis controls the magnitude of the off-diagonal elements of the Cholesky factor $\boldsymbol{L}$. The vertical axis depicts the ratio of the mean variance of the OMT estimator to that of the RT estimator for the off-diagonal elements of $\boldsymbol{L}$.

range of the model hyperparameter α_0 considered. Note that as we crank up the shape augmentation setting B, the RSVI variance approaches that of the pathwise gradient.[12]

7.1.3. Multivariate Normal

In Fig. 5 we use synthetic test functions to illustrate the amount of variance reduction that can be achieved with the OMT gradient estimator for the multivariate Normal distribution. The dimension is $D = 50$; the results are qualitatively similar for different dimensions.

[12] As discussed in Sec. 6, the variance of the RSVI gradient estimator can also be reduced by dropping the score function-like term (at the cost of some bias).

7.2. Real World Datasets

In this section we investigate the performance of our gradient estimators for the Gamma, Beta, and multivariate Normal distributions in two variational inference tasks on real world datasets. Note that we include an additional experiment for the multivariate Normal distribution in Sec. 10 of the supplementary materials. All the experiments in this section were implemented in the Pyro[13] probabilistic programming language.

7.2.1. Sparse Gamma DEF

The Sparse Gamma DEF (Ranganath et al., 2015) is a probabilistic model with multiple layers of local latent random variables $z_{nk}^{(\ell)}$ and global random weights $w_{kk'}^{(\ell)}$ that mimics the architecture of a deep neural network. Here each n corresponds to an observed data point x_n, ℓ indexes the layer, and k and k' run over the latent components. We consider Poisson-distributed observations x_{nd} for each dimension d. Concretely, the model is specified as[14]

$$z_{nk}^{(\ell)} \sim \text{Gamma}\left(\alpha_z, \frac{\alpha_z}{\sum_{k'} z_{nk'}^{(\ell+1)} w_{k'k}^{(\ell)}}\right) \quad \ell = 1, ..., L-1$$

$$x_{nd} \sim \text{Poisson}\left(\sum_{k'} z_{nk'}^{(1)} w_{k'd}^{(0)}\right) \quad z_{nk}^{L} \sim \text{Gamma}\left(\alpha_z, \alpha_z\right)$$

We set $\alpha_z = 0.1$ and use $L = 3$ layers with 100, 40, and 15 latent factors per data point (for $\ell = 1, 2, 3$, respectively). We consider two model variants that differ in the prior placed on the weights. In the first variant we place Gamma priors over the weights with $\alpha = 0.3$ and $\beta = 0.1$. In the second variant we place β' priors over the weights with the same means and variances as in the first variant.[15] The dataset we consider is the Olivetti faces dataset,[16] which consists of 64×64 grayscale images of human faces. In Fig. 6 we depict how the training set ELBO increases during the course of optimization. We find that on this task the performance of the OMT gradient estimator is nearly identical to RSVI.[17] Figure 6 suggests that gradient variance is not the limiting factor for this particular task and dataset.

[13] http://pyro.ai

[14] Note that this experiment closely follows the setup in (Ruiz et al., 2016) and (Naesseth et al., 2017).

[15] If $z \sim \text{Beta}(\alpha, \beta)$ then $\frac{z}{1-z} \sim \beta'(\alpha, \beta)$. Thus like the Gamma distribution the Beta prime distribution has support on the positive real line.

[16] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

[17] Note that we do not compare to any alternative estimators such as G-REP, since (Naesseth et al., 2017) shows that RSVI has superior performance on this task.

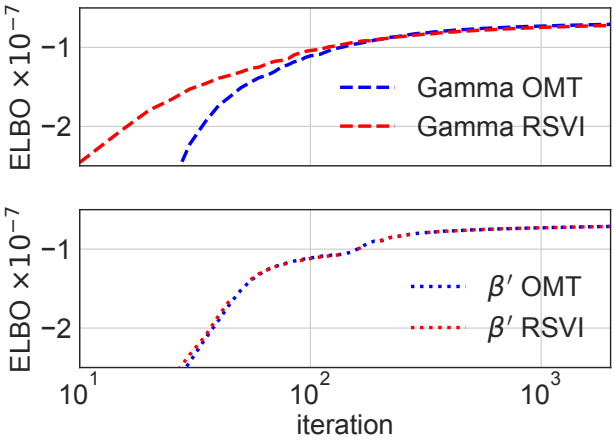

Figure 6. ELBO during training for two variants of the Sparse Gamma DEF, one with and one without Beta random variables. We compare the OMT gradient to RSVI. At each iteration we depict a multi-sample estimate of the ELBO with $N = 100$ samples.

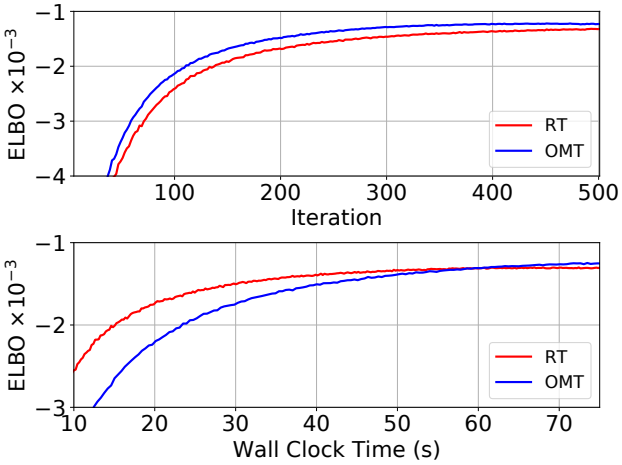

Figure 7. ELBO during training for the Gaussian Process regression task in Sec. 7.2.2. At each iteration we depict a multi-sample estimate of the ELBO with $N = 100$ samples. We compare the OMT gradient estimator to the RT estimator.

7.2.2. Gaussian Process Regression

In this section we investigate the performance of our OMT gradient for the multivariate Normal distribution, Eqn. 19, in the context of a Gaussian Process regression task. We model the Mauna Loa CO_2 data from (Keeling & Whorf, 2004) considered in (Rasmussen, 2004). We use a structured kernel that accommodates a long term linear trend as well as a periodic component. We fit the GP using a single-sample Monte Carlo ELBO gradient estimator and all $D = 468$ data points. The variational family is a multivariate Normal distribution with a Cholesky parameterization for the covariance matrix. Progress on the ELBO during the course of training is depicted in Fig. 7. We can see that the OMT gradient estimator has superior sample efficiency due to its lower variance. By iteration 270 the OMT gradient estimator has attained the same ELBO that the RT estimator attains at iteration 500. Since each iteration of the OMT estimator is ~ 1.9x slower than the corresponding RT iteration, the superior sample efficiency of the OMT estimator is largely canceled when judged by wall clock time. Nevertheless, the lower variance of the OMT estimator results in a higher ELBO than that obtained by the RT estimator.

8. Discussion and Future Work

We have seen that optimal transport offers a fruitful perspective on pathwise gradients. On the one hand it has helped us formulate pathwise gradients in situations where this was assumed to be impractical. On the other hand it has focused our attention on a particular notion of optimality, which led us to develop a new gradient estimator for the multivariate Normal distribution. A better understanding of this notion of optimality and, more broadly, a better understanding of when pathwise gradients are preferable over score function gradients (or vice versa) would be useful in guiding the practical application of these methods.

Since each solution of the transport equation Eqn. 12 yields an unbiased gradient estimator, the difference between any two such estimators can be thought of as a control variate. In the case of the multivariate Normal distribution, where computing the OMT gradient has a cost $\mathcal{O}(D^3)$, an attractive alternative to using $\boldsymbol{v}^{\text{OMT}}$ is to adaptively choose $\boldsymbol{v}$ during the course of optimization in direct analogy to adaptive control variate techniques. In future work we will explore this approach in detail, which promises lower variance than the OMT estimator at reduced computational cost.

The geometric picture from optimal transport—and thus the potential for non-trivial derivative applications—is especially rich for multivariate distributions. Here we have explored the multivariate Normal and Dirichlet distributions in some detail, but this just scratches the surface of multivariate distributions. It would be of general interest to develop pathwise gradients for a broader class of multivariate distributions, including for example mixture distributions. Rich distributions with low variance gradient estimators are of special interest in the context of SGVI, where the need to approximate complex posteriors demands rich families of distributions that lend themselves to stochastic optimization. In future work we intend to explore this connection further.

Acknowledgements

We thank Peter Dayan and Zoubin Ghahramani for feedback on a draft manuscript and other colleagues at Uber AI Labs—especially Noah Goodman and Theofanis Karaletsos— for stimulating conversations during the course of this work. We also thank Christian Naesseth for clarifying details of the experimental setup for the deep exponential family experiment in (Naesseth et al., 2017).

References

Ambrosio, Luigi, Gigli, Nicola, and Savaré, Giuseppe. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

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

Casella, George and Robert, Christian P. Rao-blackwellisation of sampling schemes. *Biometrika*, 83 (1):81–94, 1996.

Figurnov, Michael, Mohamed, Shakir, and Mnih, Andriy. Implicit reparameterization gradients. *arXiv preprint arXiv:1805.08498*, 2018.

Glasserman, Paul. *Monte Carlo methods in financial engineering*, volume 53. Springer Science & Business Media, 2013.

Glynn, Peter W. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10): 75–84, 1990.

Grathwohl, Will, Choi, Dami, Wu, Yuhuai, Roeder, Geoff, and Duvenaud, David. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. *arXiv preprint arXiv:1711.00123*, 2017.

Graves, Alex. Stochastic backpropagation through mixture density distributions. *arXiv preprint arXiv:1607.05690*, 2016.

Hoffman, Matthew D, Blei, David M, Wang, Chong, and Paisley, John. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Jordan, Michael I, Ghahramani, Zoubin, Jaakkola, Tommi S, and Saul, Lawrence K. An introduction to variational methods for graphical models. *Machine learning*, 37(2): 183–233, 1999.

Keeling, Charles David and Whorf, Timothy P. Atmospheric co2 concentrations derived from flask air samples at sites in the sio network. *Trends: a compendium of data on Global Change*, 2004.

Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Knowles, David A. Stochastic gradient variational bayes for gamma approximating distributions. *arXiv preprint arXiv:1509.01631*, 2015.

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.

Miller, Andrew, Foti, Nick, D'Amour, Alexander, and Adams, Ryan P. Reducing reparameterization gradient variance. In *Advances in Neural Information Processing Systems*, pp. 3711–3721, 2017.

Mnih, Andriy and Gregor, Karol. Neural variational inference and learning in belief networks. *arXiv preprint arXiv:1402.0030*, 2014.

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.

Paisley, John, Blei, David, and Jordan, Michael. Variational bayesian inference with stochastic search. *arXiv preprint arXiv:1206.6430*, 2012.

Paszke, Adam, Gross, Sam, Chintala, Soumith, Chanan, Gregory, Yang, Edward, DeVito, Zachary, Lin, Zeming, Desmaison, Alban, Antiga, Luca, and Lerer, Adam. Automatic differentiation in pytorch. 2017.

Price, Robert. A useful theorem for nonlinear devices having gaussian inputs. *IRE Transactions on Information Theory*, 4(2):69–72, 1958.

Ranganath, Rajesh, Gerrish, Sean, and Blei, David. Black box variational inference. In *Artificial Intelligence and Statistics*, pp. 814–822, 2014.

Ranganath, Rajesh, Tang, Linpeng, Charlin, Laurent, and Blei, David. Deep exponential families. In *Artificial Intelligence and Statistics*, pp. 762–771, 2015.

Rasmussen, Carl Edward. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pp. 63–71. Springer, 2004.

Rezende, Danilo Jimenez, Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Robbins, Herbert and Monro, Sutton. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Ross, Sheldon M. *Simulation*. Academic Press, San Diego, 2006.

Ruiz, Francisco R, AUEB, Michalis Titsias RC, and Blei, David. The generalized reparameterization gradient. In *Advances in Neural Information Processing Systems*, pp. 460–468, 2016.

Salimans, Tim, Knowles, David A, et al. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.

Schulman, John, Heess, Nicolas, Weber, Theophane, and Abbeel, Pieter. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems*, pp. 3528–3536, 2015.

Titsias, Michalis and Lázaro-Gredilla, Miguel. Doubly stochastic variational bayes for non-conjugate inference. In *International Conference on Machine Learning*, pp. 1971–1979, 2014.

Tucker, George, Mnih, Andriy, Maddison, Chris J, Lawson, John, and Sohl-Dickstein, Jascha. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. In *Advances in Neural Information Processing Systems*, pp. 2624–2633, 2017.

Villani, Cédric. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Wilks, S.S. *Mathematical Statistics*. John Wiley and Sons Inc., 1962.

Williams, Ronald J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Wingate, David and Weber, Theophane. Automated variational inference in probabilistic programming. *arXiv preprint arXiv:1301.1299*, 2013.