Pathwise Derivatives Beyond the Reparameterization Trick

Terentyev Alexander

31 октября 2024 г.

Содержание

Introduction

Motivation

Maximizing objective functions via gradient methods is ubiquitous in machine learning. Computing exact gradients w.r.t. the parameters is often unfeasible so that optimization methods must instead make due with stochastic gradient estimates. However, the gradient estimator exhibits large variance, stochastic optimization algorithms may be impractically slow. We use this perspective to compute (approximate) pathwise gradients for probability distributions not directly amenable to the reparameterization trick: Gamma, Beta, and Dirichlet

Stochastic Gradient Variational Inference

SGV Examples

One area where stochastic gradient estimators play a particularly central role is stochastic variational inference. This is especially the case for black-box methods, where conjugacy and other simplifying structural assumptions are unavailable, with the consequence that Monte Carlo estimators become necessary.

ELBO

Let $p(\mathbf{x}, \mathbf{z})$ define a joint probability distribution over observed data x and latent random variables z. One of the main tasks in Bayesian inference is to compute the posterior distribution $p(z|x) = \frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{x})}$.

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

Score Function Estimator

REINFORCE

The score function estimator, also referred to as the log derivative trick or REINFORCE, provides a simple and broadly applicable recipe for estimating ELBO gradients.

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

where $\log r = \log p(\mathbf{x}, \mathbf{z}) \log q(\mathbf{z})$.

Disadvantages

Although the score function estimator is very general it typically suffers from high variance, although this can be mitigated with the use of variance reduction techniques such as Rao-Blackwellization and control variates.

Reparameterization trick

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.

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

Disadvantages

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 $\mathcal{T}(\epsilon;\theta)$ inevitably involve special functions.

Univariate distributions

Univariate Pathwise Gradients

Consider an objective function given as

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

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

Idea

A natural choice is to use the standard uniform distribution $\mathcal{U}$,

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

Unfortunately, for many continuous univariate distributions of interest (e.g. the Gamma and Beta distributions) the transformation F_θ^{-1} (as well as its derivative w.r.t. 0) does not admit a simple analytic expression.

Implicit differentiation

Idea

Fortunately, by making use of implicit differentiation we can compute the gradient in without explicitly introducing F_θ^{-1}

$$u \equiv F_\theta(z) = \int_{-\infty}^{z} q_\theta(z')\, dz'.$$

Differentiate both sides of equation w.r.t.θ

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

Master formula

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

Univariate Pathwise Gradient

Objective function

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{q_\theta(z)} \left[\frac{df_\theta(z)}{dz} \frac{dz}{d\theta} + \frac{\partial f_\theta(z)}{\partial \theta} \right], \quad \textit{where } \frac{dz}{d\theta} = -\frac{\frac{\partial F_\theta}{\partial \theta}(z)}{q_\theta(z)}$$

Necessary function

While this derivation is elementary, it helps to clarify things: the key ingredient needed to compute pathwise gradients in the equation is the ability to compute (or approximate) the derivative of the CDF, i.e. $\frac{\partial}{\partial \theta} F_\theta(z)$

Multivariate Pathwise Gradients

The Transport Equation

Consider a multivariate distribution $q_\theta(\mathbf{z})$ in D dimensions. As we vary θ we move $q_\theta(\mathbf{z})$ along a curve in the space of distributions over the sample space. This intuitive picture can be formalized with the transport equation:

$$\frac{\partial}{\partial\theta}(q_\theta) + \nabla_{\mathbf{z}} \cdot (q_\theta v_\theta) = 0,$$

where velocity field v_θ is a vector field defined on the sample space that displaces samples $\mathbf{z}$ as we vary θ infinitesimally.

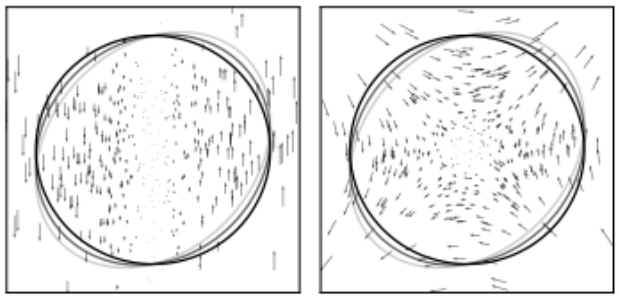

Рис.: Velocity field

Loss Function

The Transport Equation

Given a solution to the equation, we can form the gradient estimator

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

Tangent Fields

To determine a unique solution—the tangent field from the theory of optimal transport—we require that

$$\frac{\partial v_i^{\mathrm{OMT}}}{\partial z_j} = \frac{\partial v_j^{\mathrm{OMT}}}{\partial z_i} \quad \forall i, j.$$

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

$$K(v) = \frac{1}{2} \int dz \, q_\theta(z) \|v\|^2$$

About optimal solution

OMT is not optimal

The ||v||2term that appears in Eqn. 15 might lead one to hope that v OMT provides gradients that minimize gradient variance. Unfortunately, the situation is more complicated. Denoting the (mean) gradient by $\mathbf{g} = \mathbb{E}_{q_0(z)}[v \cdot \nabla_z f] - \|g\|^2$ the total gradient variance is given by

$$\mathbb{E}_{q_0(z)}\left[\|v \cdot \nabla_z f\|^2\right] - \|g\|^2$$

OMT is approximation of optimum

Still, for many choices of $f(\mathbf{z})$ 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 $f(\mathbf{z})$)).

Beta

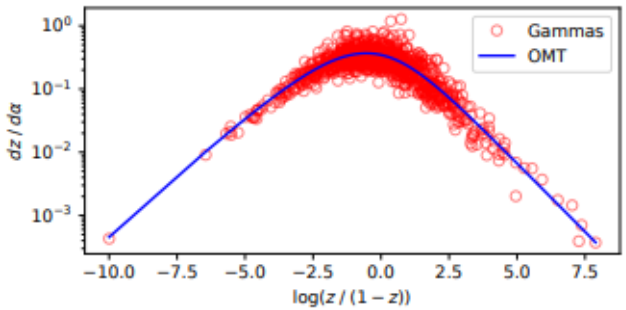

Рис.: Derivatives $\frac{dz}{d\alpha}$ for samples $z \sim \text{Beta}(1, 1)$.

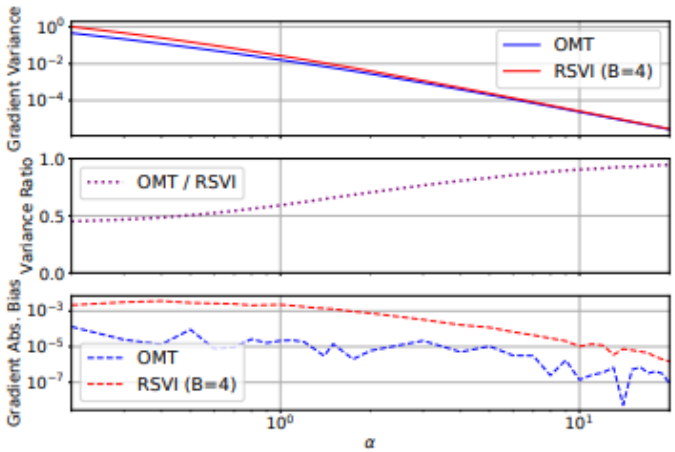

Рис.: We compare the OMT gradient to the RSVI gradient with $B = 4$ for the test function $f(z) = z^3$

Dirichlet

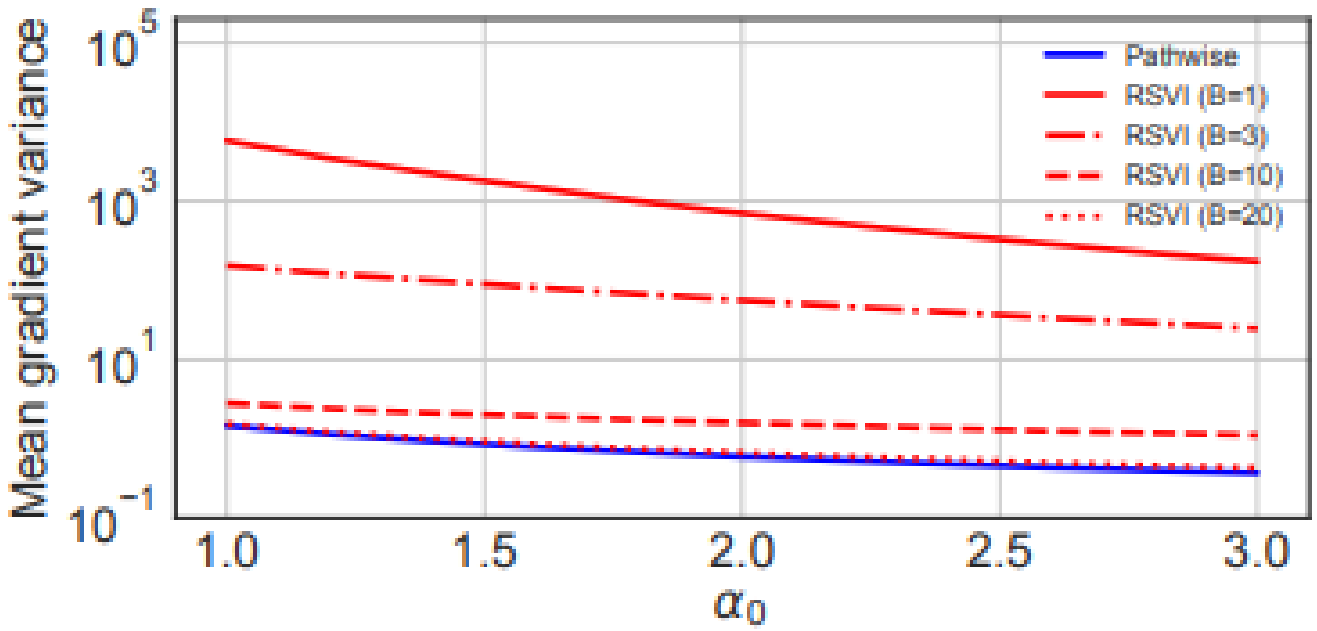

Рис.: . Gradient variance for the ELBO of a conjugate Multinomial-Dirichlet model. We compare the pathwise gradient to RSVI for different boosts B. S

Quetions

TThat's all