Stochastic Gradient Estimation
With Finite Differences

Lars Buesing, Theophane Weber, Shakir Mohamed
DeepMind
{l Louiseing, theophane, shakir}google.com

Abstract
Choosing the parameters of a probability distribution in order to minimize an ex-
pected loss is the central problem in many machine learning applications, including
inference in latent variable models or policy search in reinforcement learning. In
most cases however, the exact gradient of the expected loss is not available as
a closed-form expression. This has led to the development of flexible gradient
estimators, with a focus on the conflicting objectives of being both general-purpose
and low-variance. If the loss is non-differentiable, as is the case in reinforcement
learning, or if the distribution is discrete, as for probabilistic models with discrete
latent variables, we have to resort to score-function (SF) gradient estimators. Naive
SF estimators have high variance and therefore require sophisticated variance re-
duction techniques, such as baseline models, to render them effective in practice.
Here we show that under certain symmetry and parametric assumptions on the
distribution, one can derive unbiased stochastic gradient estimators based on finite
differences (FD) of the loss function. These estimators do not require learning
baseline models and potentially have less variance. Furthermore, we highlight
connections of the FD estimators to simultaneous perturbation sensitivity analysis
(SPSA), as well as weak derivative and "straight-through" gradient estimators.

1 Introduction
Numerous problems in engineering, operations research, and machine learning naturally take the
form of minimizing an expected loss:
\[ \min_\theta \mathbb{E}_{x \sim P_\theta}[H(x)]. \] (1)
Here we allow parameters to stochastically influence the loss function \( H \) through the random variable
\( x \). Often, one can neither compute exact expectations of \( H \) wrt. \( P_\theta \), nor its gradients \( \partial_\theta \mathbb{E}[H] \); in this
case one has to rely on simulation-based / stochastic optimization by evaluating the loss on samples
of \( x \). Nevertheless, one can still apply approximate gradient-based optimization techniques using
the so-called score-function (SF) estimator. SF estimators are central to policy-gradient methods
in RL and have also been applied to learning probabilistic models with discrete latent variables.

As the SF estimator applies to very general optimization problems, it is not surprising that in practice it
is dogged by high variance, necessitating intricate variance reduction techniques. Here we show, that
if we are willing to assume additional properties about how the parameters influence the distribution
\( P_\theta \), we can make use of this additional information to arrive at a estimator which leverages more
structure of the problem. More specifically, we assume that the density of \( P_\theta \) is an even function wrt.
\( x \) and the parameters we are optimizing are “location-scale” parameters. If, furthermore, it is possible
to evaluate the loss function multiple times on the same input, we show that the standard SF estimator
can be replaced by an unbiased estimator based on finite-differences (FD) of the loss function. This

\footnote{For ease of notation we suppress the deterministic dependence of \( H(x, \theta) \) and simply write \( H(x) \). The
results extend to the general case.}

Submitted to NIPS 2016 workshop “Advances in Approximate Bayesian Inference”. Do not distribute.
estimator uses approximate first and second order derivative information and therefore potentially renders the optimization much more tractable. The requirement of multiple evaluations of $H$ excludes some general problems (such as general RL problems), but includes (potentially challenging) special cases such as bandit problems.

2 Finite difference estimators

2.1 Score-function estimator

For optimizing the criterion $E[H]$ by gradient-descent methods, we can use an estimator $g_{\theta}$ that approximates the gradient of the cost in expectation $\partial_{\theta}E[H] \approx E[g_{\theta}]$. A very general, unbiased gradient estimator is the score-function (SF) estimator $g_{\theta, SF}$:

$$g_{\theta, SF}(x, \theta) = H(x)\partial_{\theta} \log f_{\theta}(x),$$

where $f_{\theta}$ is the probability density of $P_{\theta}$.

2.2 Symmetric location-scale distributions

Assume that $P_{\theta}$ for $\theta = (\mu, \sigma) \in \mathbb{R}^d \times \mathbb{R}_+^d$ is a “location-scale” family of distributions with mean parameter $\mu$ and scale parameter $\sigma$ and independent components:

$$f_{\theta}(x) = \prod_{i=1}^{d} \frac{1}{\sigma_i} f_i \left( \frac{x_i - \mu_i}{\sigma_i} \right).$$

Here, $f_i$ are the densities which we require to be even functions:

$$f_i(-\epsilon_i) = f_i(\epsilon_i).$$

Under these assumptions we can express the gradient of the expected loss wrt. the location parameter $\mu$ in the following way:

$$\partial_{\mu} E[H] = E[H(x)\partial_{\mu} \log f_{\theta}(x)]$$
$$= -\sigma_i^{-1}E_{\epsilon \sim P_{(0,1)}} [H(\mu + \sigma \epsilon) s_i(\epsilon_i)]$$
$$= -\frac{1}{2} \sigma_i^{-1}E_{\epsilon \sim P_{(0,1)}} [(H(\mu + \sigma \epsilon) - H(\mu - \sigma \epsilon)) s_i(\epsilon_i)],$$

where $s_i$ is the score function:

$$s_i(\epsilon_i) := \frac{f_i'(\epsilon_i)}{f_i(\epsilon_i)}.$$

In the derivation we critically used the fact that the score function $s_i$ is odd for an even density $f_i$, i.e. $s_i(-\epsilon_i) = -s_i(\epsilon_i)$. A similar reasoning can be applied for deriving the gradients wrt. the scale parameter $\sigma$, leading to the gradient estimators:

$$g_{\mu, FD}(\epsilon, \theta) = -\frac{1}{2} \sigma^{-1} s(\epsilon) (H(\mu + \sigma \epsilon) - H(\mu - \sigma \epsilon))$$
$$g_{\sigma, FD}(\epsilon, \theta) = -\frac{1}{2} \sigma^{-1} (s(\epsilon) \epsilon + 1) (H(\mu + \sigma \epsilon) - 2H(\mu) + H(\mu - \sigma \epsilon))$$

where $0, 1$ are vectors of 0s and 1s of appropriate size, $\epsilon := (\epsilon_1, \ldots, \epsilon_d)^T$ is defined analogously, and $\sigma^{-1}$ as well as $s(\epsilon) \epsilon$ are to be interpreted elementwise. We call the above quantities finite difference (FD) estimators, as they involve finite differences of first and second order of the cost $H$ for mean and scale parameters respectively. It is straightforward to show that they are unbiased. Furthermore, they are valid estimators even for non-differentiable cost functions $H$. 
where

\[ \hat{L} = E_{x \sim P_q}[H(x, \phi)] + \mu^T E[g_{\mu,FD}(\epsilon, \bar{\mu})] + \sigma^T E[g_{\sigma,FD}(\epsilon, \bar{\sigma})]. \]

Here, we set \( \hat{\theta} := (\hat{\mu}, \sigma) \) := \( \theta \) during the forward pass (ie. function evaluation), but we do not back-propagate gradients wrt. \( \hat{\theta} \) in the backward-pass. We also allow for the loss \( H \) to depend on additional parameters \( \phi \) (which can also be functions of \( \theta \)); if \( H \) is differentiable wrt. parameters \( \phi \), they can also be learned by gradient descent methods.

**Lotka-Volterra ODEs** Consider the Lotka-Volterra (LV) system of ODEs for modeling population sizes \( a, b \) of the prey / predator species:

\[
\frac{da}{dt} = \alpha a - \beta ab, \quad \frac{db}{dt} = \delta ab - \gamma b.
\]

Assume we are given a set of \( N \) noisy observations \( \{o^n_{0:T}\}, n = 1, \ldots, N \) with \( o^n_0 := (a^n_0, b^n_0) \) consisting of trajectories of length \( T \) and initial conditions \( o^n_0 \). We wish to approximate the posterior distribution over the model parameters \( \tilde{x} = (\alpha, \beta, \gamma, \delta) \) with a log-normal distribution \( q \), i.e. \( \tilde{x} = \exp(x) \) and \( q(x) = \mathcal{N}(x|\mu, \sigma) \). We do so by optimizing the variational lower bound on the marginal likelihood:

\[
L(\mu, \sigma) := E_{x \sim q} \left[ \frac{1}{2} \sum_{n=1}^{N} \|o^n_{1:T} - \hat{o}(o^n_0, x)\|^2 + \log p(x) - \log q(x) \right],
\]

where \( \hat{o}(o^n_0, x) \) is the output of an LV simulator with initial conditions \( o^n_0 \) and parameters \( \exp(x) \). Comparing the results of optimizing \( L \) using different gradient estimators shown in fig. 1, we can see that FD estimators outperform naive SF and SF with baseline estimators. Methods were compared based on the same number of function calls to the LV simulator.

**Discussion.** The FD estimators can be interpreted as applying the standard variance reduction technique of antithetic sampling to the SF estimator. In antithetic sampling (assuming an even density \( f(\theta) \)) one uses samples \( \epsilon \) and their negation \( -\epsilon \) to estimate expected values. This makes use of the fact that the expectation of an odd function under an even distribution is 0. The FD estimator \( s(\epsilon)(H(\mu + \sigma \epsilon) - H(\mu - \sigma \epsilon)) \) is exactly the even part of the SF estimator. We therefore expect the former to have lower variance compared to the latter if \( H \) is roughly an odd function around \( \mu \).
Perturbation Analysis. Evaluating the FD estimators requires three function calls to $H$ with inputs $\mu + \sigma \epsilon_i$, $\mu$ and $\mu - \sigma \epsilon_i$; hence they are roughly three times as expensive as SF estimators. However, in contrast to the latter, they do not require a learnable baseline model, which is essential to reduce variance of regular SF estimators in practice. The above FD estimators exhibit an interesting parallel to SPSA [9]: a naive finite difference estimator for the $d$-dimensional gradient would require $2d$ evaluations of the form $H(\mu + \Delta \epsilon_i)$, where $\epsilon_i$ is the $i$-th basis vector. SPSA is based on a simultaneous perturbation of all coordinates, i.e. $H(\mu + \Delta) - H(\mu - \Delta)$; empirically this leads to a $d$-fold speedup over the naive FD implementation. The FD gradient estimators [2] and [3] are also based on simultaneous perturbations and we therefore expect them to inherit the speedup of SPSA compared to independent perturbations. In contrast to SPSA, the statistics of the perturbations do not have to be chosen, but are given by the current estimate of the scale parameters $\sigma$.

Straight-through estimators. Let us assume that $H$ is differentiable and that values of $\sigma_i \epsilon_i$ are typically small. By definition of the gradient, we have

$$H(\mu + \sigma e) - H(\mu - \sigma e) \approx 2 \sum_i \sigma_i \epsilon_i \partial_{\mu_i} H(\mu),$$

which turns the FD estimator into

$$E[g_{\mu_i,FD}] = -\partial_{\mu_i} H(\mu) E[\epsilon_i f'_i(\epsilon_i)] = \partial_{\mu_i} H(\mu),$$

which is the straight through estimator (where $E[\epsilon_i f'_i(\epsilon_i)] = -1$ by integration by parts) introduced by [3].

Weak derivative estimators. The proposed FD estimators as intimately connected to the concept of weak derivatives (WD) of distributions [6]. Briefly, a 3-tuple $(c(\theta), \partial f^+, \partial f^-)$ consisting of a $\theta$-dependent constant $c$ and two probability distribution $\partial f^\pm$ is called a weak derivative of $P_\theta$ if:

$$\partial_\theta E_{x \sim P_\theta}[H(x)] = c(\theta) (E_{x \sim \partial f^+}[H(x)] - E_{x \sim \partial f^-}[H(x)]).$$

For the one-dimensional case, making use of the assumptions that $f_0(x)$ is an even function of $x$, it is straight-forward to show that:

$$g_{\mu,WD} \propto H(\mu + \Delta) - H(\mu - \Delta); \quad \Delta \sim \partial f^+.$$

The support of $\partial f^+$ is $\mathbb{R}^+$ and $\partial f^+(x) \propto s(x)f_0(x)$. So, in 1-d for a location parameter, we can interpret the FD estimator [2] approximating the WD estimator by importance sampling with a proposal density $f_0$ and the resulting importance weight $s(x)$. For the case of $d > 1$ the WD estimator is computationally expensive as it requires $2d$ evaluations of $H(\mu \pm \Delta \epsilon_i + \Delta_{-i})$, where $\Delta_i \sim \partial f^+$ and $(\Delta_{-i})_j \sim f$ for $j \neq i$ and $(\Delta_{-i})_i = 0$.

References


