Taylor Residual Estimators via Automatic Differentiation

Andrew C. Miller *
Harvard University
Cambridge, MA 02138
acm@seas.harvard.edu

Nicholas J. Foti
University of Washington
Seattle, WA 98195
nfoti@uw.edu

Ryan P. Adams
Princeton, Google Brain
Princeton, NJ 08544
rpa@princeton.edu

Abstract

We develop a method for reducing the variance of Monte Carlo estimators against distributions with known moments, termed Taylor residual Monte Carlo estimators (TREs). We analyze the variance of TREs, and derive conditions under which TREs outperform the original Monte Carlo estimators in terms of estimator variance. Additionally, modern automatic differentiation tools can be leveraged to efficiently compute these new estimators. The utility of TREs is demonstrated on a Monte Carlo variational inference problem.

1 Introduction

Many fundamental problems in machine learning and statistics can be framed as the expectation of a function of a random variable. For example, modern variational inference algorithms for complex probabilistic models hinge on well-behaved Monte Carlo estimates of gradients. If the variance of the estimated gradient is large then gradient-based optimization can exhibit chaotic behavior or require such small step-sizes that the algorithm does not converge in a reasonable amount of time. A common approach is a Monte Carlo estimator, where the random variable is sampled (perhaps multiple times), the function is computed, and the values are averaged. The variance of the Monte Carlo estimate is a crucial property when applying Monte Carlo methods since a large variance can make the estimate unreliable. There has been a large body of literature on controlling the variance of Monte Carlo estimates such as *control variates* that reduce variance using a correlated estimate with the same mean as the original estimate. However, obtaining variance reduction is still a challenging problem.

In this work we develop a family of Monte Carlo estimators based on the Taylor expansion of the function being integrated. These estimators can be efficient to compute and easy to implement with modern automatic differentiation tools. We can interpret the resulting estimator as a control variate and we study the conditions under which the variance of the estimator is reduced. We apply the estimator to a Monte Carlo variational inference problem and show that the method achieves lower variance estimates of gradients.

2 Taylor Residual Monte Carlo Estimator

Let $X \in \mathbb{R}^D$ be a random variable with distribution π. Consider a function $f : \mathbb{R}^D \mapsto \mathbb{R}$ whose expectation we would like to take with respect to π which we write $\mathbb{E}_\pi [f] = \int f(x)\pi(dx)$. In this work we assume the we can easily draw i.i.d. samples from π. The standard Monte Carlo estimator of $\mathbb{E}_\pi[f]$ is constructed by sampling from π and then computing the sample mean of f:

$$x^{(n)} \sim \pi, \quad \hat{f} = \frac{1}{N} \sum_{n=1}^{N} f(x^{(n)}). \tag{1}$$

*http://andymiller.github.io

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

While Eq. (1) is an extremely general way to estimate an expectation, it can be inefficient to ignore known structure in f and π which can manifest as a large amount of variance in the $\hat{f}$. We will assume that all moments of π are known and computable and we denote the mth moment of π about the point x_0 as

$$\mathcal{M}_{x_0}^{(m)} = \int (x - x_0)^m \pi(dx) \,. \tag{2}$$

Now, consider decomposing f into (i) its first order Taylor expansion around x_0 and (ii) the residual:

$$f(x) = \underbrace{f(x_0) + (x - x_0)^\intercal \left[\frac{\partial f}{\partial x}(x_0)\right]}_{\triangleq f_{x_0}^{(1)}(x)} + R_{x_0}^{(1)}(x), \tag{3}$$

where the Taylor remainder $R_{x_0}^{(1)}(x)$ can be determined from the second-order derivatives of f. We can re-write the target expectation as

$$\mathbb{E}_\pi[f] = \mathbb{E}_\pi\left[f_{x_0}^{(1)}(x) + R_{x_0}^{(1)}(x)\right] = f(x_0) + \mathbb{E}_\pi[(x - x_0)]^\intercal \left[\frac{\partial f}{\partial x}(x_0)\right] + \mathbb{E}_\pi\left[R_{x_0}^{(1)}(x)\right] \tag{4}$$

$$= f(x_0) + \mathcal{M}_{x_0}^{(1)^\intercal} \left[\frac{\partial f}{\partial x}(x_0)\right] + \mathbb{E}_\pi\left[R_{x_0}^{(1)}(x)\right], \tag{5}$$

In general, we can use an M^{th}-order Taylor expansion about x_0 and write the expectation as

$$\mathbb{E}_\pi[f] = f(x_0) + \sum_m \mathcal{M}_{x_0}^{(m)} \left[\frac{\partial^m f}{\partial x^m}(x_0)\right] + \mathbb{E}_\pi\left[R_{x_0}^{(M)}(x)\right]. \tag{6}$$

In this case the Taylor remainder $R_{x_0}^{(M)}(x)$ can be found from the $(M+1)^{\text{st}}$ order derivatives of f.

Since we assume all moments $\mathcal{M}_{x_0}^{(m)}$ are known we see that all of the randomness in the estimators given in Eqs. (3) and (6) comes from the expectation of the remainder term. We call estimators of this form *Taylor residual Monte Carlo estimators* (TREs). Note that we have simply shifted the variance of the Monte Carlo estimate into the higher-order derivatives of the function. As such, we can expect the residual to have low variance when the low order derivatives well-approximate f around x_0. Taylor residual Monte Carlo estimates can be viewed as performing approximate Rao-Blackwellization in that the aspects of f captured in the low-order derivatives is being integrated out and replaced with non-random quantities. Furthermore, using modern automatic differentiation tools [Maclaurin et al., 2015, Abadi et al., 2016, core team, 2017] we can easily compute higher order derivatives of scalar functions and the requisite tensor contractions.

We can interpret the first-order Taylor residual estimate in Eq. (3) as a control-variate estimator, implying that TREs may achieve smaller variance than that of pure Monte Carlo estimators. To see the connection to control variates, consider a single sample first order TRE:

$$\mathbb{E}_\pi[f] = f(x_0) + \mathcal{M}_{x_0}^{(1)^\intercal} \left[\frac{\partial f}{\partial x}(x_0)\right] + R_{x_0}^{(1)}(x) \tag{7}$$

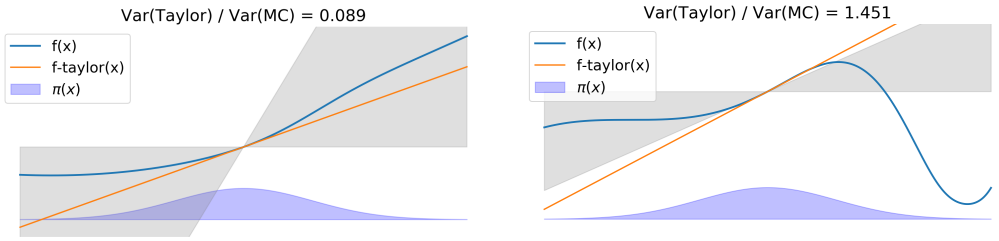
$$= f(x_0) + \mathcal{M}_{x_0}^{(1)^\intercal} \left[\frac{\partial f}{\partial x}(x_0)\right] + \left[f(x) - \left(f(x_0) + (x - x_0)^\intercal \left[\frac{\partial f}{\partial x}(x_0)\right]\right)\right] \tag{8}$$

$$= f(x) - \left(\mathcal{M}_{x_0}^{(1)} - (x - x_0)\right)^\intercal \left[\frac{\partial f}{\partial x}(x_0)\right] \tag{9}$$

where we recognize Eq. (9) as the equation for a control variate with scale coefficient 1. In fact, first-order Taylor residual estimators generalize the reduced variance gradient estimators presented in Miller et al. [2017] and provide a framework to study when such gradient estimators will be effective. In the next section we study the variance properties of TREs to determine conditions under which we attain variance reduction.

3 Variance Analysis

The Taylor residual estimator is useful if its variance is lower than that of the standard Monte Carlo estimator. In this section we will show that the variance properties of the TRE depend on the



(a) Variance reduced

(b) Variance increased

Figure 1: Illustration of the conditions for TRE variance reduction. In each example, the gray area indicates the set of linear approximations to $f(x)$ that result in decreased variance, as indicated by Equation (19). (a) When the first-order Taylor approximation of f at x_0 (orange line) is in the gray region then the corresponding TRE will have smaller variance than the Monte Carlo estimator. Functions that are close to linear in the range of π will have a larger region where variance reduction occurs while highly nonlinear functions will have smaller regions. (b) The TRE estimator can have larger variance than the MC estimator when the gradient at x_0 falls outside of the gray region.

relationship between the locally linear Taylor approximation and the global linear structure captured by linear least squares regression.

Consider the Monte Carlo estimator given in Eq. (11) and the first order Taylor residual estimator in Eq. (12) using a single sample from π. For notational simplicity, we take $x_0 = 0$ and define $f_0 = f(0)$, as well as $f_0' = \frac{\partial f}{\partial x}(0)$ as shorthand. We write the estimators as

$$x \sim \pi \qquad \text{sample from distribution} \tag{10}$$

$$\hat{f} = f(x) \qquad \text{Monte Carlo estimator} \tag{11}$$

$$\hat{f}_1 = f(x) - (f_1(x) - \mathbb{E}[f^{(1)}]) \qquad \text{First order Taylor residual estimator} \tag{12}$$

$$= f(x) - xf_0' + \mu f_0', \tag{13}$$

where $\mu = \mathbb{E}(x)$ is the known first moment of $\pi(x)$. The variances of the two estimators are then

$$\mathbb{V}(\hat{f}) = \mathbb{E}\left[\hat{f}^2\right] - \mathbb{E}[f]^2 \tag{14}$$

$$\mathbb{V}(\hat{f}_1) = \mathbb{V}(f(x) - xf_0' + \mu f_0') = \mathbb{E}\left[(f(x) - xf_0')^2\right] - \left(\mathbb{E}[f] - \mu f_0'\right)^2. \tag{15}$$

We want to find sufficient conditions such that the variance of the new estimator is smaller than the original, $\mathbb{V}(\hat{f}) \geq \mathbb{V}(\hat{f}_1)$. We first substitute the variances with their definitions into the inequality

$$\mathbb{E}[f(x)^2] - \mathbb{E}[f]^2 \geq \mathbb{E}\left[(f(x) - xf_0')^2\right] - \left(\mathbb{E}[f] - \mu f_0'\right)^2. \tag{16}$$

Expanding the two quadratics, and canceling terms, we get

$$0 \geq \mathbb{E}[x^2](f_0')^2 - 2f_0'\mathbb{E}[xf(x)] - \mu^2(f_0')^2 + 2f_0'\mu\mathbb{E}[f] \tag{17}$$

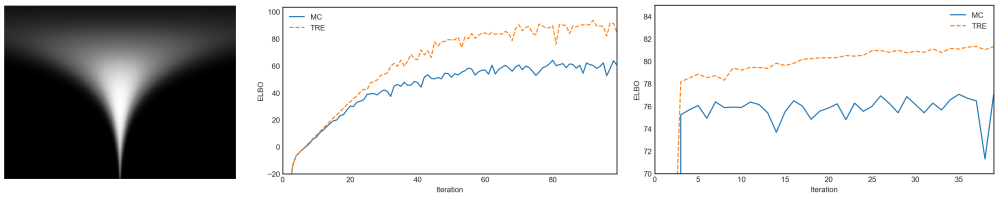
$$= (f_0')^2\mathbb{V}(x) - 2f_0'\mathbb{E}[xf(x)] + 2f_0'\mu\mathbb{E}[f] \tag{18}$$

$$\implies 1 \leq \frac{2}{f_0'}\frac{\mathbb{C}(x, f(x))}{\mathbb{V}(x)} \implies |f_0'| \leq 2\left|\frac{\mathbb{C}(x, f(x))}{\mathbb{V}(x)}\right|. \tag{19}$$

Eq. (19) indicates a relationship between linear control-variate methods and linear least-squares regression. Since $\mathbb{V}(x)^{-1}\mathbb{C}(x, f(x))$ is the population least squares solution for f regressed on x, we see that variance reduction depends on whether the first order Taylor expansion of f is within a cone around the linear least squares approximation. We visually depict both successful and unsuccessful variance reduction for a one-dimensional example in Fig. 1. Appendix A further explores this bound.

4 Experiments

We demonstrate the variance reduction capabilities of TREs in the context of Monte Carlo variational inference. Specifically, we compare a TRE to the pure Monte Carlo estimator on the variational



(a) Target distribution (b) Gaussian Approximation (c) Normalizing Flows Approximation (4 layers)

Figure 2: Comparision of Taylor residual and Monte Carlo estimators on Monte Carlo variational inference optimization using both a Gaussian variational distribution and a normalizing flow. In both cases, TREs provide lower variance gradient estimates and attain higher lower-bounds.

evidence lower bound (ELBO). We target a 20-dimensional "funnel" distribution that exhibits features typical of posteriors from hierarchical models Neal [2003]. A bivariate marginal of the "funnel" is depicted in Fig. 2a. We consider two variational approximations, a Gaussian and a normalizing flow distribution and show that TREs attain lower variance estimates and yield more robust optimization.

Gaussian Approximation For variational approximation $q(x; \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{\lambda}_\mu, \boldsymbol{\lambda}_\sigma)$, with variational parameters $\boldsymbol{\lambda}$, the Monte Carlo ELBO estimator can be computed by first drawing a sample $x \sim q(x; \boldsymbol{\lambda})$, and then computing

$$f(x) = \ln \pi(x, \mathcal{D}) - \ln q(x; \boldsymbol{\lambda}). \tag{20}$$

We optimize the ELBO by using estimators of the gradient of Eq (20) with respect to $\boldsymbol{\lambda}$. We compute the pathwise gradient estimator (reparameterization gradient) [Glasserman, 2004] for both the MC and TRE estimators, and use these noisy gradient estimates in gradient ascent.

At a random initialization of $\boldsymbol{\lambda}$, we measure the variance of the first order Taylor residual estimator to be about 320 times lower than the Monte Carlo estimator (for 2 samples). We show the results of ELBO optimization in Figure 2 using a 2-sample Monte Carlo estimator and a 2-sample Taylor residual estimator. The TRE estimator has a smaller variance for more iterations than the MC estimator allowing it to attain larger ELBO values. After convergence, we measure the TR estimator to have .8 the variance of the MC estimator.

Normalizing Flows We also apply the Taylor residual estimator to a more flexible posterior approximation, a planar normalizing flow distribution [Rezende and Mohamed, 2015]. A normalizing flow distributed random variable is constructed by applying a sequence of parameterized invertible maps to a simple random variable (e.g. $x_0 \sim \mathcal{N}(0, I_D)$). Here, we broke the ELBO into two pieces

$$\mathcal{L}(\boldsymbol{\lambda}) = \underbrace{\mathbb{E}_q[\ln \pi(x, \mathcal{D})]}_{\text{model term}} - \underbrace{\mathbb{E}[\ln q(x; \boldsymbol{\lambda})]}_{\text{entropy term}}. \tag{21}$$

Unlike for the Gaussian variational family where the entropy term can be computed exactly and the model term is the only random component, for normalizing flows, we must estimate the entropy term using Monte Carlo. Here, we apply a TRE to the model term and use the simple Monte Carlo estimator for the entropy term. We found this resulted in consistent variance reduction compared to the Monte Carlo estimator. At initialization we measure a $40\times$ variance reduction over the standard Monte Carlo estimator, and a $2\times$ reduction at convergence. Fig. 2c shows the results of optimization using the TRE where it is clear that the optimization is more stable.

5 Conclusion

We presented Taylor residual estimators to efficiently compute lower variance Monte Carlo estimators by using a Taylor expansion. We showed that when a selected locally linear Taylor approximation aligns with the global least squares linear approximation the proposed estimator will have lower variance than the standard Monte Carlo estimator. The advantages of the TRE method were demonstrated on performing Monte Carlo variational inference where we obtained more robust optimization results under two variational approximations. We plan to extend the method to estimate highly nonlinear functions using a hierarchical approach that combines locally linear approximations.

Acknowledgments

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

PyTorch core team. Pytorch, 2017. URL `http://pytorch.org/`.

Paul Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53. Springer Science & Business Media, 2004.

Dougal Maclaurin, David Duvenaud, Matthew Johnson, and Ryan P. Adams. Autograd: Reverse-mode differentiation of native Python, 2015. URL `http://github.com/HIPS/autograd`.

Andrew C Miller, Nicholas J Foti, Alexander D'Amour, and Ryan P Adams. Reducing reparameterization gradient variance. In *Advances in Neural Information Processing Systems*, 2017.

Radford M Neal. Slice sampling. *Annals of statistics*, pages 705–741, 2003.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1530–1538, 2015.

A Variance Analysis

Now the question is, under what conditions of $f(x)$ and $\pi(x)$ is this condition true? We can start by re-writing the covariance using a taylor-expanded $f(x)$

$$\mathbb{C}(x, f(x)) = \mathbb{E}[xf(x)] - \mathbb{E}[x]\mathbb{E}[f(x)] \tag{22}$$

$$= \mathbb{E}\left[x(f_0 + \sum_{n=1}^{\infty} \frac{1}{n!} f_0^{(n)} x^n)\right] - \mu\mathbb{E}\left[f_0 + \sum_{n=1}^{\infty} \frac{1}{n!} f_0^{(n)} x^n\right] \tag{23}$$

$$= \sum_{n=1} \frac{1}{n!} f_0^{(n)} \left(\mathbb{E}[x^{n+1}] - \mu\mathbb{E}[x^n]\right) \tag{24}$$

$$= \sum_{n=1} \frac{1}{n!} f_0^{(n)} \left(\mathcal{M}_\pi^{(n+1)} - \mu\mathcal{M}_\pi^{(n)}\right) \qquad \text{moments of } \pi \tag{25}$$

The first term of this series is a simple function of the variance of π

$$\sum_{n=1} \frac{1}{n!} f_0^{(n)} \left(\mathcal{M}_\pi^{(n+1)} - \mu\mathcal{M}_\pi^{(n)}\right) \tag{26}$$

$$= f_0^{(1)}(\mathcal{M}_\pi^{(2)} - \mu\mathcal{M}_\pi^{(1)}) + \sum_{n=2} \frac{1}{n!} f_0^{(n)} \left(\mathcal{M}_\pi^{(n+1)} - \mu\mathcal{M}_\pi^{(n)}\right) \tag{27}$$

$$= f_0^{(1)}\mathbb{V}(x) + \sum_{n=2} \frac{1}{n!} f_0^{(n)} \left(\mathcal{M}_\pi^{(n+1)} - \mu\mathcal{M}_\pi^{(n)}\right) \tag{28}$$

So we can express the inequality above as a bound on the variance of $x \sim \pi$ as a function of the higher moments of π and derivatives of f

$$\sum_{n=2} \frac{1}{n!} \frac{f_0^{(n)}}{f_0^{(1)}} \left(\mathcal{M}_\pi^{(n+1)} - \mu\mathcal{M}_\pi^{(n)}\right) \leq \frac{1}{2}\mathbb{V}(x) \tag{29}$$

For the first order estimator to have reduced variance, a scaled sum of the difference of higher order moments needs to be smaller than the variance of $x \sim \pi$.

For example, if $\pi(x) = \mathcal{N}(0, \sigma^2)$, then the n'th even moment is $\sigma^n(n-1)!!$ (note that $(n-1)!!$ is the double factorial, which is the product of a decreasing sequence of numbers with the same parity, e.g. $(n-1)(n-3)(n-5)...$), and the odd moments are zero. We can write the inequality as

$$\sum_{n=2} \frac{1}{n!} \frac{f_0^{(n)}}{f_0^{(1)}} \sigma^{(n+1)}(n)!! = \sum_{n=2} \frac{f_0^{(n)}}{f_0^{(1)}} \frac{\sigma^{(n+1)}}{(n-1)!!} \leq \frac{1}{2}\mathbb{V}(x) \tag{30}$$

So in this case we can see that the inequality is easily achieved when the variance of π, σ^2, is small, and when the ratio of higher order derivatives to the first derivative $\frac{f_0^{(n)}}{f_0^{(1)}}$ is small.