Taylor Residual Estimators via Automatic Differentiation

HARVARD
School of Engineering and Applied Sciences

Andrew C. Miller [1] Nicholas J. Foti [2] Ryan P. Adams [3]

[1]Harvard [2]University Washington [3]Princeton and Google Brain

Summary

- Monte Carlo estimators lie at the heart of many algorithms, e.g. importance sampling, variational inference, generative adversarial networks.
- Variance of Monte Carlo estimators can have a profound effect on algorithmic efficiency and robustness.
- When the sampling distribution is tractable, we can leverage differentiable structure to construct new estimators.
- We introduce *Taylor Residual Estimators* that leverage known moments to attain lower variance than the naive Monte Carlo estimator.
- We show how to construct these estimators using automatic differentiation, analyze their variance, and apply them to a variational inference problem.

Taylor Residual Estimators

- Let $X \in \mathbb{R}^D$ be a random variable with distribution π with ***known moments***.
- We want to estimate $\mathbb{E}_\pi[f] = \int f(x)\pi(dx)$, for function $f : \mathbb{R}^D \mapsto \mathbb{R}$
- Standard Monte Carlo estimator: sample from π and compute sample mean:

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

- *Can be inefficient to ignore known structure in f and π.*
- Denote the mth moment of π about point x_0 as

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

- Decompose f into (i) 1st-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 remainder $R_{x_0}^{(1)}(x)$ is a function of higher-order derivatives of f.
- We can re-write the target expectation as

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

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

$$\mathbb{E}_\pi[f] = f(x_0) + \sum_m \mathcal{M}_{x_0}^{(m)}\left[\frac{\partial^m f}{\partial x^m}(x_0)\right] + \mathbb{E}_\pi\left[R_{x_0}^{(M)}(x)\right]. \tag{4}$$

In this case the Taylor remainder $R_{x_0}^{(M)}(x)$ can be found from the $(M+1)^{\text{st}}$ order derivatives of f.
- The randomness in the estimators in Eqs. (3) and (4) comes from the *remainder term*. This is reminiscent of *Rao-Blackwellization*.
- ***Taylor residual estimators* (TREs) estimate these remaindor terms.**
- TREs can also be interpreted as control-variate estimators:

$$\mathbb{E}_\pi[f] = f(x_0) + \mathcal{M}_{x_0}^{(1)\intercal}\left[\frac{\partial f}{\partial x}(x_0)\right] + R_{x_0}^{(1)}(x) \tag{5}$$
$$= f(x) - \left(\mathcal{M}_{x_0}^{(1)} - (x - x_0)\right)^\intercal \left[\frac{\partial f}{\partial x}(x_0)\right] \tag{6}$$

Variance Analysis

Big Question: When does the TRE have lower variance?

- Recall the MC and first order Taylor Estimators where we define $x_0 = 0$, $f_0 = f(0)$, and $f_0' = \frac{\partial f}{\partial x}(0)$:

$x \sim \pi$	sample from distribution	(7)
$\hat{f} = f(x)$	Monte Carlo estimator	(8)
$\hat{f}_1 = f(x) - (f_1(x) - \mathbb{E}[f^{(1)}])$	First order Taylor residual estimator	(9)
$= f(x) - xf_0' + \mu f_0'$,		(10)

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

$$\mathbb{V}(\hat{f}) = \mathbb{E}\left[\hat{f}^2\right] - \mathbb{E}[f]^2 \tag{11}$$
$$\mathbb{V}(\hat{f}_1) = \mathbb{V}(f(x) - xf_0' + \mu f_0') = \mathbb{E}\left[(f(x) - xf_0')^2\right] - (\mathbb{E}[f] - \mu f_0')^2\,. \tag{12}$$

- We find conditions sucht that

$$\underbrace{\mathbb{V}(\hat{f})}_{\text{MC}} \geq \underbrace{\mathbb{V}(\hat{f}_1)}_{\text{TRE}}.$$

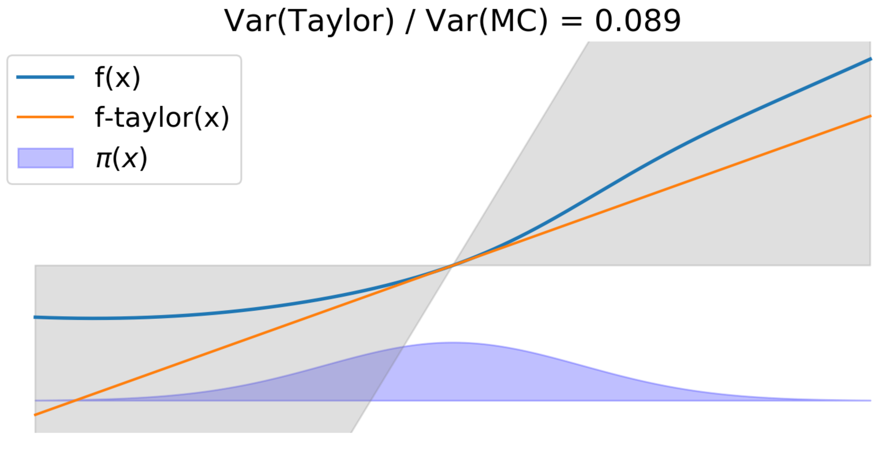
- First, substitute the variances with their definitions into the inequality

$$\mathbb{E}[f(x)^2] - \mathbb{E}[f]^2 \geq \mathbb{E}\left[(f(x) - xf_0')^2\right] - (\mathbb{E}[f] - \mu f_0')^2\,. \tag{13}$$

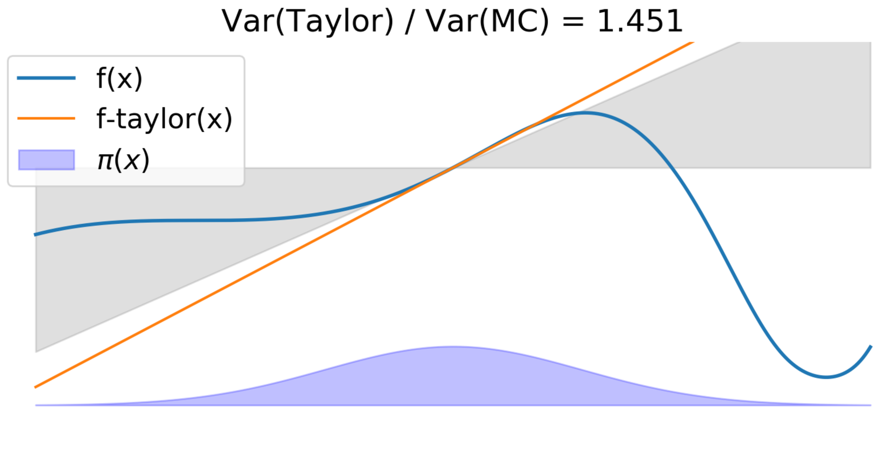
Expanding the two quadratics, and canceling terms, we get

$$0 \geq \mathbb{E}[x^2](f_0')^2 - 2f_0'\mathbb{E}[xf(x)] - \mu^2(f_0')^2 + 2f_0'\mu\mathbb{E}[f] \tag{14}$$
$$= (f_0')^2\mathbb{V}(x) - 2f_0'\mathbb{E}[xf(x)] + 2f_0'\mu\mathbb{E}[f] \tag{15}$$
$$\implies 1 \leq \frac{2}{f_0'}\frac{\mathbb{C}(x, f(x))}{\mathbb{V}(x)} \implies |f_0'| \leq 2\left|\frac{\mathbb{C}(x, f(x))}{\mathbb{V}(x)}\right|. \tag{16}$$

- Eq. (15) indicates a relationship between linear control-variate methods and linear least-squares regression: $\mathbb{V}(x)^{-1}\mathbb{C}(x, f(x))$ is the population least squares solution for f regressed on x.
- **Variance reduction depends on whether the first order Taylor expansion of f is within a cone around the linear least squares approximation.**



(a) TRE has lower variance than MC estimator.



(b) TRE has higher variance than MC estimator.

Figure: 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 (15). (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

Example

Setting:
Monte Carlo VI for a "Funnel" posterior distribution.

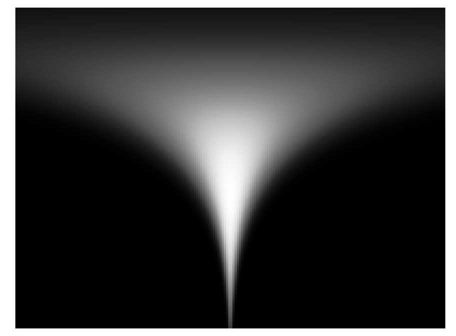


Figure: Funnel target distribution

- The variational objective is the ELBO, which we approximate with Monte Carlo by drawing a sample $x \sim q(x; \boldsymbol{\lambda})$, and then computing

$$f(x) = \ln \pi(x, \mathcal{D}) - \ln q(x; \boldsymbol{\lambda}). \tag{17}$$

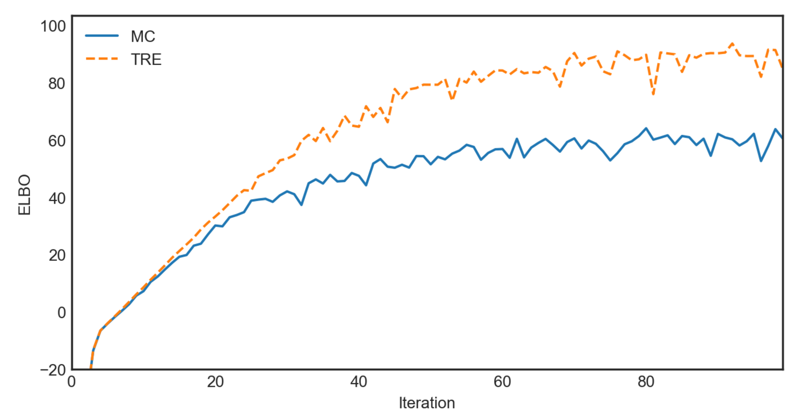
- We use TREs to fit an approximation from two different variational families: diagonal Gaussians and Normalizing Flows.

Gaussian Approximation:
- Define variational approximation $q(x; \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{\lambda}_\mu, \boldsymbol{\lambda}_\sigma)$, with parameters $\boldsymbol{\lambda}$.
- We optimize the ELBO by using estimators of the gradient of Eq (16) with respect to $\boldsymbol{\lambda}$.
- We compute the pathwise gradient estimator (reparameterization gradient) (**?**) 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 2-sample Monte Carlo estimator.
- We show the results of ELBO optimization in Figure 3 using a 2-sample Monte Carlo estimator and a 2-sample TRE. The TRE has a smaller variance for more iterations than the MC estimator allowing it to attain larger ELBO values for the step-size. After convergence, we measure the TRE to have .8 the variance of the MC estimator.

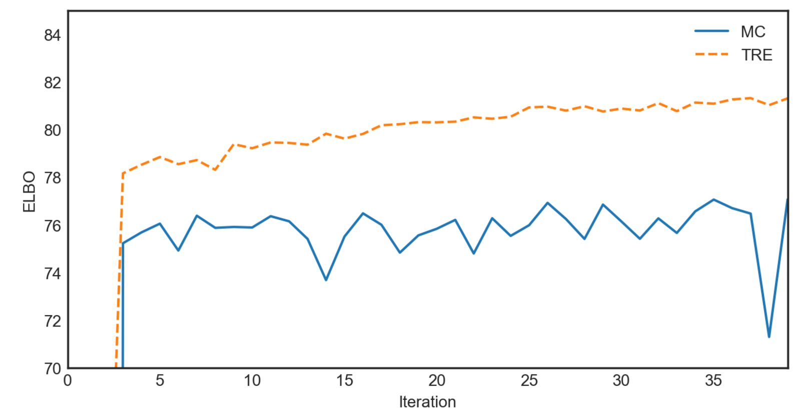
Normalizing Flows: We also apply the Taylor residual estimator to a more flexible posterior approximation, a planar normalizing flow distribution (**?**).
- We broke the ELBO into two pieces:

$$\mathcal{L}(\boldsymbol{\lambda}) = \underbrace{\mathbb{E}_q[\ln \pi(x, \mathcal{D})]}_{\text{model term}} - \underbrace{\mathbb{E}[\ln q(x; \boldsymbol{\lambda})]}_{\text{entropy term}}. \tag{18}$$

- Unlike for the Gaussian variational family, where the entropy term can be computed exactly, 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. 3b shows the results of optimization using the TRE where it is clear that the optimization is more stable.



(a) Gaussian Approximation



(b) Normalizing Flows Approximation (4 layers)

Figure: 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.

acm@seas.harvard.edu