Comparing LS and EVM in Variance Reduction

Spoiler: Empirical Risk Minimization beats Least Squares and enjoys fast convergence rates

Leonid Iosipoi
Higher School of Economics
Faculty of Computer Science
iosipoileonid@gmail.com
http://hdilab.ru/

Introduction

Suppose that we wish to estimate $\mathsf{E}[f(X)]$, where X is a random vector in $\mathcal{X} \subset \mathbb{R}^d$ with a density $\pi(x)$ and $f : \mathbb{R}^d \to \mathbb{R}$ with $\mathsf{Var}[f(X)] < \infty$.

If the dimension d is large or/and the function f is complicated, the only acceptable computational way is Monte Carlo method

$$\mathsf{E}[f(X)] \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i),$$

where $X_1, \dots, X_n$ is an independent sample from $\pi(x)$.

The Monte Carlo estimator has an error variance of the form

$$\mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right) = \frac{\sigma^2}{n}, \quad \text{where } \sigma^2 = \mathsf{Var} f(X).$$

We can make the variance smaller by using a larger value of n, but the cost of the corresponding estimate also grows with n. Therefore, it is important to find a way to reduce σ^2 instead of increasing the sample size n. Methods to reduce the variance σ^2 are called Variance Reduction Techniques.

Control Variates method

The method of control variates is a popular technique in Monte Carlo integration that aims at reducing the variance of the naive Monte Carlo estimate. The idea is the following.

1. Fix a class G of functions $g(x)$ with $\mathsf{E}[g(X)] = 0$;
2. Find $g^* \in G$ such that the variance of $f(X) - g^*(X)$ is minimized.

The aim is to reduce the variance

$$\inf_{g \in G} \mathsf{Var}[f(X) - g(X)] \ll \mathsf{Var}[f(X)].$$

The problem is how to solve this optimization problem numerically, as usually the variance of $f(X) - g(X)$ can not be computed analytically. There are two general approaches to estimate g^*; namely, empirical variance minimization (EVM) and least squares (LS) approach.

- $\widehat{g}_{\text{evm}} = \operatorname{arginf}_{g \in \mathcal{G}} Q_{\text{evm}}(g)$, where $Q_{\text{evm}}(g) = V_n(f - g)$; here V_n denotes the sample variance,
- $\widehat{g}_{\text{ls}} = \operatorname{arginf}_{g \in \mathcal{G}} Q_{\text{ls}}(g)$, where $Q_{\text{ls}}(g) = \sum_{i=1}^{n}(f(X_i) - g(X_i))^2$.

In the literature the LS method is more popular since its analysis is simpler, see, for instance, [Oates et al. (2016)], [Portier and Segers (2018)], and references therein. On the contrary, we are not aware of any theoretical or empirical studies of the EVM method. Note that the global minima of expected values of Q_{evm} and Q_{ls} coincide. This fact makes us expect similar performance of both approaches. But it is not the case as we see in the next section.

Illustrative examples

A popular way to construct a class of control variates G is to use the Stein's identity. Namely,

$$g_{\mathbf{a}}(x) = \sum_{i=1}^{d} \frac{1}{\pi(x)} \frac{\partial}{\partial x_i}\big((a_{i0} + a_{i1}x_i + a_{i2}x_i^2 + a_{i3}x_i^3)\pi(x)\big), \quad \mathbf{a} \in \mathbb{R}^{d \times 4}.$$

Under reasonably weak conditions, it holds $\mathsf{E}[g_{\mathbf{a}}(X)] = 0$ for all $\mathbf{a} \in \mathbb{R}^{d \times 4}$.

Our numerical experiments are organized as follows. We generate a training sample of size n_{train} to compute $\widehat{\mathbf{a}}_{\text{evm}}$ and $\widehat{\mathbf{a}}_{\text{ls}}$ by minimization Q_{evm} and Q_{ls} correspondingly. Then we generate a testing sample of size n_{test} to compute the sample variance $\widehat{\sigma}^2 = V_n(f)$. The same testing sample is used to compute the sample variances $\widehat{\sigma}^2_{\text{evm}} = V_n(f - g_{\widehat{\mathbf{a}}_{\text{evm}}})$, $\widehat{\sigma}^2_{\text{ls}} = V_n(f - g_{\widehat{\mathbf{a}}_{\text{ls}}})$. We also compute efficiencies as $\text{eff}_{\text{evm}} = \widehat{\sigma}^2/\widehat{\sigma}^2_{\text{evm}}$ and $\text{eff}_{\text{ls}} = \widehat{\sigma}^2/\widehat{\sigma}^2_{\text{ls}}$. The results for different functions f and densities are presented in Table 1, Table 2, and Table 3.

Discussion. In one-dimensional case the EVM method performs a little bit better than the LS method but there are no clear leader. On the contrary, Table 2 tells us that the EVM method performs significantly better than the LS method in multivariate setting. We see that even for a simple function as $f(x) = \|x\|^2$, the original sample variance $\widehat{\sigma}^2$ is less than the "reduced" sample variance $\widehat{\sigma}^2_{\text{ls}}$ obtained by the LS method on a test sample. It means that the LS method is, in a sense, unstable when dimension of a problem grows. An interesting observation is that the EVM method performs well even on small training samples, see Table 3.

Table 1: Comparison of EVM and LR approaches in **1-dimensional** case with a training sample of size $\mathbf{n_{train} = 500}$ and a test sample of size $\mathbf{n_{test} = 100\,000}$. We consider Normal distribution $\mathcal{N}(0,1)$ and Exponential distribution $\text{Exp}(1)$.

Distribution & Function	$\widehat{\sigma}^2$	$\widehat{\sigma}^2_{\text{evm}}$	$\widehat{\sigma}^2_{\text{ls}}$	eff_{evm}	eff_{ls}
Normal distribution					
$f(x) = x^2$	1.9989	$3.2 \cdot 10^{-15}$	0.0064	$6.0 \cdot 10^{14}$	312.3281
$f(x) = e^x$	4.6410	0.0272	0.0319	170.4984	145.0551
$f(x) = \cos(x)$	0.1999	0.0008	0.0016	248.3388	119.2143
$f(x) = 1/(1 + \lvert x\rvert)$	0.0346	0.0105	0.0087	3.2844	3.9780
Exponential distribution					
$f(x) = x^2$	19.9852	$3.0 \cdot 10^{-13}$	0.0042	$6.3 \cdot 10^{13}$	4660.4100
$f(x) = \cos(x)$	0.3492	0.0431	0.0422	8.1006	8.2629
$f(x) = 1/(1 + \lvert x\rvert)$	0.0479	0.0012	0.0017	39.4416	26.9634

Table 2: Comparison of EVM and LR approaches in **10-dimensional** case with a training sample of size $\mathbf{n_{train} = 500}$ and a test sample of size $\mathbf{n_{test} = 100\,000}$. We consider random vectors with independent components distributed according to Normal distribution $\mathcal{N}(0,1)$ and Exponential distribution $\text{Exp}(1)$.

Distribution & Function	$\widehat{\sigma}^2$	$\widehat{\sigma}^2_{\text{evm}}$	$\widehat{\sigma}^2_{\text{ls}}$	eff_{evm}	eff_{ls}
Normal distribution					
$f(x) = \|x\|^2$	20.0487	$1.0 \cdot 10^{-13}$	37.7377	$1.9 \cdot 10^{14}$	0.5310
$f(x) = \sum_i e^{x_i}$	46.1526	0.3992	104.6210	115.5993	0.4410
$f(x) = \sum_i \cos(x_i)$	2.0038	0.0102	13.5536	194.7966	0.1476
$f(x) = 1/(1 + \|x\|)$	0.0020	0.0003	0.0246	6.8433	0.0834
Exponential distribution					
$f(x) = \|x\|^2$	193.939	$1.4 \cdot 10^{-12}$	1461.3000	$1.3 \cdot 10^{14}$	0.0442
$f(x) = \sum_i \cos(x_i)$	3.4982	2.2988	73.6570	1.5216	0.0158
$f(x) = 1/(1 + \|x\|)$	0.0031	0.0007	0.1542	4.4274	0.0204

Table 3: Comparison of EVM and LR approaches in **10-dimensional** case with a training sample of size $\mathbf{n_{train} = 50}$ and a test sample of size $\mathbf{n_{test} = 100\,000}$. We consider random vectors with independent Standard normal $\mathcal{N}(0,1)$ components.

Distribution & Function	$\widehat{\sigma}^2$	$\widehat{\sigma}^2_{\text{evm}}$	$\widehat{\sigma}^2_{\text{ls}}$	eff_{evm}	eff_{ls}
Normal distribution					
$f(x) = \|x\|^2$	20.0487	$1.5 \cdot 10^{-10}$	1508.8300	$1.2 \cdot 10^{11}$	0.0132
$f(x) = \sum_i e^{x_i}$	46.1526	1.5104	4058.0300	30.5547	0.0111
$f(x) = \sum_i \cos(x_i)$	2.0038	0.0286	557.9050	69.9258	0.0033
$f(x) = 1/(1 + \|x\|)$	0.0020	0.0048	0.9988	0.4260	0.0018

Theoretical results

In this section we show that EVM is suppose to work better than the LS approach in a general setting. First let us fix a class of control variates $g_\phi(x)$ parameterized by $\phi \in \Phi$,

$$g_\phi(x) = \frac{1}{\pi(x)} \frac{\partial^d}{\partial x_1 \dots \partial x_d}\big(\phi(x)\pi(x)\big)$$

for a smooth function $\phi : \mathcal{X} \to \mathbb{R}$. Again, under reasonably weak conditions, it holds $\mathsf{E}[g_\phi(X)] = 0$ for any $\phi \in \Phi$. Other forms of $g_\phi(x)$ can also be considered. Since now Φ is not a finite dimensional vector space we will call the functions g_ϕ control functionals.

Consider the equation $f(x) - g_\phi(x) = \mathsf{E}[f(X)]$ in ϕ. For any solution ϕ^* satisfying this equation it holds $\mathsf{Var}[f(X) - g_{\phi^*}(X)] = 0$. Conversely, $\mathsf{Var}[f(X) - g_\phi(X)] = 0$ only when ϕ satisfies the latter equation. Hence, solutions to $f(x) - g_\phi(x) = \mathsf{E}[f(X)]$ is what we want to get solving LS or EVM optimization problems.

Finally, let $C^s_{\text{poly}\uparrow}(\mathcal{X})$ be a set of functions with derivatives growing not faster than a polynomial, i.e.

$$C^{s+1}_{\text{poly}\uparrow}(\mathcal{X}) := \{\phi \in C^s(\mathcal{X}) : \ \exists m \in \mathbb{N}, \text{ s.t. } |\phi^{(k)}(x)| \lesssim |x|^m \text{ as } |x| \to \infty, \ \forall k = 0, \dots, s\}$$

and let $C^s_{\text{exp}\downarrow}(\mathcal{X})$ be a set of functions with derivatives decaying as an exponential, i.e.

$$C^s_{\text{exp}\downarrow}(\mathcal{X}) := \{h \in C^s(\mathcal{X}) : \ \exists \alpha > 0, \text{ s.t. } |h^{(k)}(x)| \lesssim e^{-\alpha} \text{ as } |x| \to \infty, \ \forall k = 0, \dots, s\}.$$

Theorem 1. *Suppose that $\pi(x) \in C^{s+1}_{exp\downarrow}(\mathcal{X})$ and $\pi(x) > 0$ on $\mathcal{X}$. Suppose also $f \in C^s_{poly\uparrow}(\mathcal{X})$. Then for $\Phi = C^{s+1}_{poly\uparrow}(\mathcal{X})$, functions leading to zero variance belong to global minima of Q_{evm} and do not belong to global minima of Q_{ls}.*

Moreover, one can take into consideration the estimation error and show that variance of $\widehat{g}_{\text{evm}}$ tends to zero as $n \to \infty$. The following proposition holds for $\widehat{g}$ a slightly different quantity to $\widehat{g}_{\text{evm}}$.

Theorem 2. *Let the assumptions of Theorem 1 hold. Fix any $1 < p < \infty$ and let $\Phi = \{\phi \in C^{s+1}_{poly\uparrow}(\mathbb{R}) : \|(\phi\pi)'\|_{W^{s,p}(\mathbb{R})} \le \|\pi(f - \mathcal{E})\|_{W^{s,p}(\mathbb{R})}\}$. Then $\phi^* \in \Phi$ and it holds with probability at least $1 - \delta$,*

$$\mathsf{Var}(f(X) - \widehat{g}(X)) \lesssim \left(\frac{1}{n}\right)^{\frac{1}{1+d/s}} + \frac{\log(\frac{1}{\delta})}{n}.$$

Conclusions

- EVM is better than LS for Variance Reduction problems, especially in multidimensional setting and small training samples.
- Despite the fact that global minima of expected values of Q_{evm} and Q_{ls} coincide, on a finite sample sizes their performance is vastly different. When dimension of a problem grows, the LS method is, in a sense, unstable.
- The variance of the EVM estimate (any global minima) tends to zero with high probability as $n \to \infty$. The rate of the excess variance is up to n^{-1}, this is the best possible rate one can achieve in general. This convergence rate is usually referred to as the fast convergence rate in the literature.

Acknowledgements

The author is greatly indebted to Denis Belomestny for suggesting the problem and for many stimulating conversations.

References

[Belomestny et al. (2018)] D. Belomestny, L. Iosipoi, and N. Zhivotovskiy. Variance Reduction via Empirical Variance Minimization. *arXiv preprint arXiv:1712.04667*, 2018.