The square root rule for adaptive importance sampling

Art B. Owen

Stanford University

Based on joint work with Yi Zhou, PhD (1998)

arXiv:1901.02976

To appear in ACM Transactions on Modeling and Simulation (TOMACS)

O & Zhou (2019)

Adaptive importance sampling

1) We use importance sampling

2) From data $\cdots$ see that we could have done it better

3) So we iterate

This talk

How to combine results from multiple iterations.
Weight k'th iteration proportionally to $\sqrt{k}$.

Simple, safe, effective.

Genesis

This is from the Appendix to

"Adaptive importance sampling by mixtures of products of beta distributions"

O & Zhou (1998)

Importance sampling notation

$$\mu = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}, \quad \boldsymbol{x}_i \overset{\mathrm{iid}}{\sim} q$$

where $q(\boldsymbol{x}) > 0$ whenever $f(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$.

Variance

$$\mathrm{var}(\hat{\mu}) = \frac{\sigma_q^2}{n}, \quad \text{where}$$

$$\sigma_q^2 = \int \frac{f^2p^2}{q} - \mu^2 = \int \frac{(fp - \mu q)^2}{q}$$

$f \geqslant 0 \implies \sigma_q = 0$ can be approached

Avoid small q

Self normalized I.S.

$$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} \Big/ \frac{1}{n}\sum_{i=1}^{n}\frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)}, \quad \boldsymbol{x}_i \overset{\text{iid}}{\sim} q$$

Less restrictive: p and q don't have to be normalized

More restrictive: we need $q > 0$ whenever $p > 0$

Nota Bene

SNIS cannot approach zero variance unless f is constant.

$$\lim_{n\to\infty} n \times \mathrm{var}(\tilde{\mu}) \geqslant \left[\int |f(\boldsymbol{x}) - \mu| p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right]^2$$

We focus here on adaptive plain IS.

Some findings apply to SNIS too.

Parametric AIS

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i) p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i; \theta)}, \quad \boldsymbol{x}_i \overset{\text{iid}}{\sim} q(\cdot; \theta)$$

Core iteration

1) choose θ,
2) get $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \to \hat{\mu}$,
3) update θ

Basic TODO list

1) pick a family $q(\cdot;\theta)$, $\theta \in \Theta$

2) choose starting point θ_1

3) choose sample size n and number $K \geqslant 2$ of steps

4) design a rule to pick θ_k using data from steps $1 \cdots k-1$

5) sample $\boldsymbol{x}_{ik} \overset{\text{iid}}{\sim} q(\cdot;\theta_k)$ and compute

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} \frac{f(\boldsymbol{x}_{ik})p(\boldsymbol{x}_{ik})}{q(\boldsymbol{x}_{ik};\theta_k)},$$

6) combine $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_K$ into $\hat{\mu}$

There are $N = nK$ data values.

This talk

is all about step 6

Example AIS

Ryu and Boyd (2014)

Adapt after every data point. $n = 1$, $K = N$, using convex optimization

Zhang (1996)

$K = 2$. First sample is a pilot sample. Second sample from a kernel density estimate.

Kollman, Baggerly, Cox, Picard (1999)

Get $\mathrm{var}(\hat{\mu}) \approx \exp(-A \times K)$. Possible because

$$f(\boldsymbol{x})p(\boldsymbol{x}) \propto q(\boldsymbol{x};\theta) \quad \text{some } \theta \in \Theta \subset \mathbb{R}^r.$$

Kong and Spanier (2011)

Geometric convergence in radiative transport problems.

De Boer, Kroese, Mannor, Rubinstein (2005)

Adaptive cross-entropy.

Martingales

History **prior** to step k: $\mathcal{H}_k \equiv (x_{i\ell},\ i = 1, \ldots, n,\ \ell < k)$

A martingale argument underlies the analysis of mean, variances, covariances.

Unbiasedness

$$\begin{aligned} & \mathbb{E}(\hat{\mu}_k \mid \mathcal{H}_k) = \mu \\ \Longrightarrow \quad & \mathbb{E}(\hat{\mu}_k) = \mathbb{E}(\mathbb{E}(\hat{\mu}_k \mid \mathcal{H}_k)) = \mu. \end{aligned}$$

Variance

$$\mathrm{var}(\hat{\mu}_k \mid \mathcal{H}_k) = \sigma_k^2 \equiv \frac{1}{n} \int \frac{(f(\boldsymbol{x})p(\boldsymbol{x}) - \mu q(\boldsymbol{x};\theta_k))^2}{q(\boldsymbol{x};\theta_k)} \,\mathrm{d}\boldsymbol{x}$$

NB $\sigma_k^2 = \sigma_k^2(\mathcal{H}_k)$ is random

$$\mathrm{var}(\hat{\mu}_k) = \mathbb{E}(\sigma_k^2) \equiv \tau_k^2$$

Variance estimates

$$\hat{\sigma}_k^2 = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\Bigl(\frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i;\theta_k)} - \hat{\mu}_k\Bigr)^2 \qquad (\text{if } n \geqslant 2)$$

$$\mathbb{E}(\hat{\sigma}_k^2 \mid \mathcal{H}_k) = \sigma_k^2$$

$$\mathbb{E}(\hat{\sigma}_k^2) = \mathbb{E}(\sigma_k^2) = \tau_k^2$$

$\hat{\sigma}_k^2$ is unbiased for **both** σ_k^2 and τ_k^2

Take $\hat{\tau}_k^2 \equiv \hat{\sigma}_k^2$

Covariance

For $\ell > k$

$$\begin{aligned}\operatorname{cov}(\hat{\mu}_k, \hat{\mu}_\ell) &= \mathbb{E}(\mathbb{E}((\hat{\mu}_k - \mu)(\hat{\mu}_\ell - \mu) \mid \mathcal{H}_\ell)) \\ &= \mathbb{E}((\hat{\mu}_k - \mu)\mathbb{E}(\hat{\mu}_\ell - \mu \mid \mathcal{H}_\ell)) \\ &= 0\end{aligned}$$

Upshot

$\hat{\mu}_k$ are unbiased and uncorrelated

Fixed linear weights

$$\hat{\mu} = \sum_{k=1}^{K} \omega_k \hat{\mu}_k \qquad \omega_k \geqslant 0 \ \text{ and } \ \sum_k \omega_k = 1$$

Variance

$$\mathrm{var}(\hat{\mu}) = \sum_{k=1}^{K} \omega_k^2 \tau_k^2$$

Unknown optimal weights

$$\omega_k \propto \tau_k^{-2}$$

Why simple?

Consider AMIS Cornuet, Marin, Mira, Robert (2012)

Weight on $\hat{\mu}_k$ can depend on future iterations. Very hard to analyze.

> "$\cdots$ the convergence properties of the algorithm cannot be investigated $\cdots$"

What not to do

Do **not** take $\omega_k \propto \hat{\tau}_k^{-2} = \hat{\sigma}_k^{-2}$

Positive skew is common

$$\begin{aligned} & \mathbb{E}((\hat{\mu}_k - \mu)^3 \mid \mathcal{H}_k) > 0 \\ \Longrightarrow\ & \mathrm{cov}(\hat{\mu}_k, \hat{\tau}_k^2) > 0 \end{aligned}$$

$\Longrightarrow$ Get small $\hat{\mu}_k$ with small $\hat{\tau}_k^2$ (large ω_k)
and large $\hat{\mu}$ with small ω_k

Result

We would downweight large $\hat{\mu}_k$ (large ω_k)
and upweight small ones
Bad for failure probabilities

Also

$\mathrm{var}(\hat{\sigma}_k^2 \mid \mathcal{H}_k) = \infty$ possible.
$\hat{\sigma}_k^2 = 0$ possible

Model for steady gain

$$\tau_k^2 = \tau^2 \times k^{-y}, \quad 0 \leqslant y \leqslant 1, \quad 0 < \tau < \infty$$

Invoking G.E.P. Box: This model might never hold exactly but it captures qualitative behavior and variance is a continuous function of the weights used.

Too pessimistic case

$y = 0 \implies$ no learning

Too optimistic case

$y = 1 \implies$ get $\mathrm{var}(\hat{\mu}) = O(N^{-2})$

Not reasonable unless $f(\boldsymbol{x})p(\boldsymbol{x}) = q(\boldsymbol{x};\theta)$ some θ

We guess $\tau_k^2 \propto k^x$

$$\hat{\mu} = \hat{\mu}(x) = \sum_{k=1}^{K} k^x \hat{\mu}_k \Big/ \sum_{k=1}^{K} k^x \quad 0 < x < 1$$

Variances

$$\tau_k^2 = \tau^2 k^{-y}$$

$$\hat{\mu}(x) = \sum_{k=1}^{K} k^x \hat{\mu}_k \Big/ \sum_{k=1}^{K} k^x$$

$$\text{var}(\hat{\mu}(x)) = \tau^2 \sum_{k=1}^{K} k^{2x-y} \Big/ \left(\sum_{k=1}^{K} k^x\right)^2$$

At $x =$ optimal unknown y

$$\text{var}(\hat{\mu}(y)) = \tau^2 \left(\sum_{k=1}^{K} k^y\right)^{-1}$$

Rate

$$\text{var}(\hat{\mu}(y)) = O(K^{-y-1}) = O(N^{-y-1})$$

Inefficiency

We should have used y but we did use x

$$\rho_K(x \mid y) \equiv \frac{\operatorname{var}(\hat{\mu}(x))}{\operatorname{var}(\hat{\mu}(y))} = \frac{\left(\sum_{k=1}^K k^{2x-y}\right)\left(\sum_{k=1}^K k^y\right)}{\left(\sum_{k=1}^K k^x\right)^2}$$

Just use $x = 1/2$

$$\sup_{1 \leqslant K < \infty} \sup_{0 \leqslant y \leqslant 1} \rho_K\left(\frac{1}{2} \mid y\right) \leqslant \frac{9}{8}$$

O & Zhou (2019)

Unknown optimal rate; mildly suboptimal constant.

Steps in the proof

Lemma 1

$$\sup_{0\leqslant y\leqslant 1} \rho_K(x \mid y) = \begin{cases} \rho_K(x \mid 1), & x \leqslant 1/2 \\ \rho_K(x \mid 0), & x \geqslant 1/2. \end{cases}$$

Trivial for $K = 1$. For $K \geqslant 2$, $\rho_K(x \mid y)$ is strictly convex in y

Also: $\rho_K\Bigl(\frac{1}{2} \mid 0\Bigr) = \rho_K\Bigl(\frac{1}{2} \mid 1\Bigr)$.

Lemma 2

$$\rho_{K+1}\Bigl(\frac{1}{2} \mid 1\Bigr) > \rho_K\Bigl(\frac{1}{2} \mid 1\Bigr), \quad K \geqslant 1$$

Long argument using very tight inequalities for sums of powers of integers.

Theorem

L'Hôpital's rule: $\lim_{K\to\infty} \rho_K\Bigl(\frac{1}{2} \mid 1\Bigr) = \frac{9}{8}$

Also

Any $x \neq 1/2$ gives some $\rho_K(x \mid y) > 9/8$.

Robustness

O & Zhou (2019) looks at other models

Diminishing returns model:

$$\tau_k^2 \propto \begin{cases} k^{-1}, & 1 \leqslant k \leqslant k_1 \\ (1+k_1)^{-1}, & k_1+1 \leqslant k \leqslant k_1+k_2 \end{cases}$$

Square root rule has

$$\max_{1\leqslant k_1\leqslant 100}\ \max_{1\leqslant k_2\leqslant 100} \rho \leqslant 1.121$$

Bad case for sqrt

First iterations make no progress.

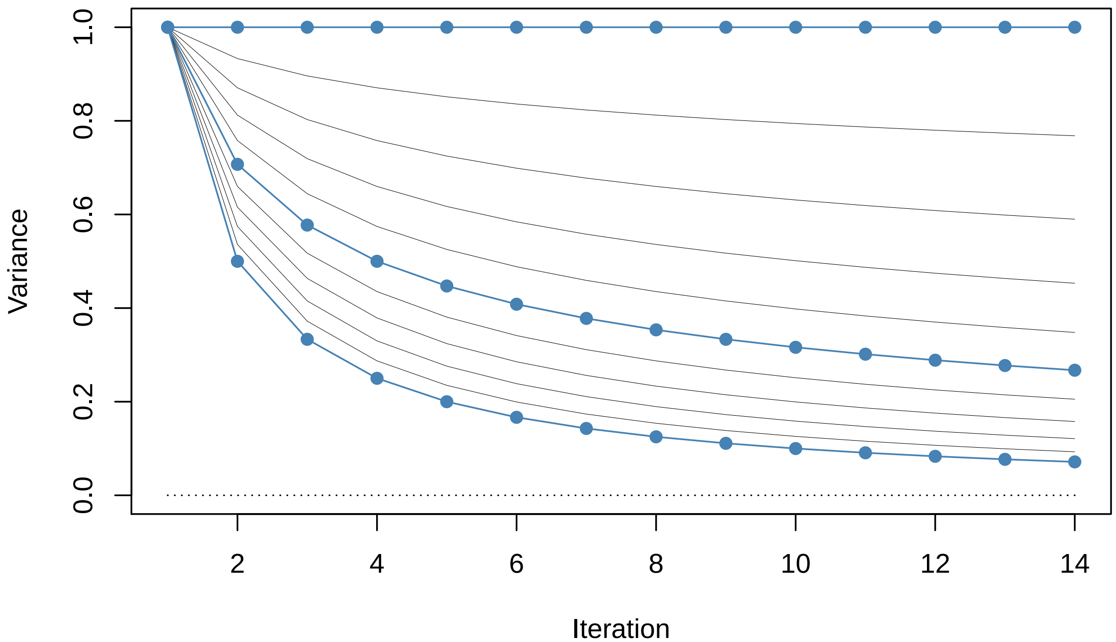
Then variance drops sharply.

Self normalized

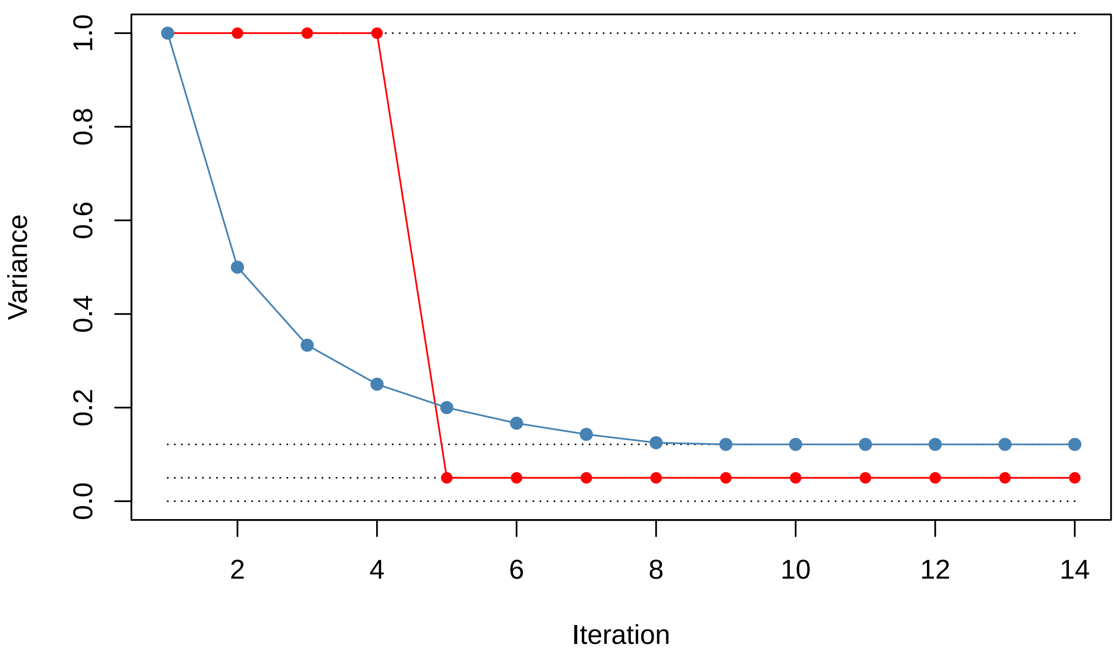
Above argument applies to variance

Have to contend with bias.

Power laws, y = 0, 0.5, 1



Asymptote => OK,   Step => inefficient



Realistic patterns

1) $\tau_k^2 \geqslant \eta > 0$ for $k = 1, \dots, K$

2) $\tau_{k+1}^2 \leqslant \tau_k^2$

3) And maybe diminishing returns

   (a) $\tau_{k+2}^2 / \tau_{k+1}^2 \geqslant \tau_{k+1}^2 / \tau_k^2$, or

   (b) $\tau_{k+1}^2 - \tau_{k+2}^2 \leqslant \tau_{k+1}^2 - \tau_k^2$

O & Zhou (2019) have some more examples.

Convex minimax

Pick ω_k in simplex to

$$\min_{\omega} \max_{\tau \in \mathcal{T}} \sum_k \omega_k^2 \tau_k^2$$

Choosing $\mathcal{T} = \{(\tau_1^2, \dots, \tau_K^2)\}$ for future work

Thanks

- Yi Zhou, co-author
- Christian Robert, Richard Everitt, invitation
- Victor Elvira, Felipe Aguayo, co-speakers
- NSF DMS-1407397, DMS-1521145, IIS-1837931
- Hobert, Betancourt, Khare, Michailidis, Patra, organizers
- Alethea Geiger, Flora Marynak, more help than we will know about