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 · · · 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

\[ \mu = \int f(\mathbf{x}) p(\mathbf{x}) \, d\mathbf{x} \]

\[ \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(\mathbf{x}_i) p(\mathbf{x}_i)}{q(\mathbf{x}_i)}, \quad \mathbf{x}_i \iid q \]

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

Variance

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

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

\( f \geq 0 \implies \sigma_q = 0 \) can be approached

Avoid small \( q \)
Self normalized I.S.

\[ \tilde{\mu} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)p(x_i)}{q(x_i)} \bigg/ \frac{1}{n} \sum_{i=1}^{n} \frac{p(x_i)}{q(x_i)}, \quad x_i \text{ iid } \sim q \]

**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.

\[ \lim_{n \to \infty} n \times \text{var}(\tilde{\mu}) \geq \left[ \int |f(x) - \mu| p(x) \, dx \right]^2 \]

We focus here on adaptive plain IS.  
Some findings apply to SNIS too.
Parametric AIS

\[ \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)p(x_i)}{q(x_i; \theta)}, \quad x_i \overset{iid}{\sim} q(\cdot; \theta) \]

Core iteration

1) choose \( \theta \),

2) get \( x_1, \ldots, x_n \), \( \to \hat{\mu} \),

3) update \( \theta \)
Basic TODO list

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

2) choose starting point $\theta_1$

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

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

5) sample $x_{ik} \sim q(\cdot; \theta_k)$ and compute

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

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

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

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

$$\mathbb{E}(\hat{\mu}_k \mid \mathcal{H}_k) = \mu$$

$$\implies \mathbb{E}(\hat{\mu}_k) = \mathbb{E}(\mathbb{E}(\hat{\mu}_k \mid \mathcal{H}_k)) = \mu.$$
Variance

\[ \text{var}(\hat{\mu}_k \mid \mathcal{H}_k) = \sigma^2_k \equiv \frac{1}{n} \int \frac{(f(x)p(x) - \mu q(x; \theta_k))^2}{q(x; \theta_k)} \, dx \]

NB \hspace{1em} \sigma^2_k = \sigma^2_k(\mathcal{H}_k) \text{ is random}

\[ \text{var}(\hat{\mu}_k) = \mathbb{E}(\sigma^2_k) \equiv \tau^2_k \]

Variance estimates

\[ \hat{\sigma}^2_k = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{f(x_i)p(x_i)}{q(x_i; \theta_k)} - \hat{\mu}_k \right)^2 \quad (\text{if } n \geq 2) \]

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

\[ \mathbb{E}(\hat{\sigma}^2_k) = \mathbb{E}(\sigma^2_k) = \tau^2_k \]

\[ \hat{\sigma}^2_k \text{ is unbiased for both } \sigma^2_k \text{ and } \tau^2_k \]

Take \[ \hat{\tau}^2_k \equiv \hat{\sigma}^2_k \]
Covariance

For $\ell > k$

\[
\text{cov}(\hat{\mu}_k, \hat{\mu}_\ell) = \mathbb{E}(\mathbb{E}((\hat{\mu}_k - \mu)(\hat{\mu}_\ell - \mu) \mid \mathcal{H}_\ell)) \\
= \mathbb{E}((\hat{\mu}_k - \mu)\mathbb{E}(\hat{\mu}_\ell - \mu \mid \mathcal{H}_\ell)) \\
= 0
\]

Upshot

$\hat{\mu}_k$ are unbiased and uncorrelated
Fixed linear weights

\[ \hat{\mu} = \sum_{k=1}^{K} \omega_k \hat{\mu}_k \quad \omega_k \geq 0 \quad \text{and} \quad \sum_k \omega_k = 1 \]

Variance

\[ \text{var}(\hat{\mu}) = \sum_{k=1}^{K} \omega_k^2 \tau_k^2 \]

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.

“··· the convergence properties of the algorithm cannot be investigated ···”
What not to do

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

Positive skew is common

$$\mathbb{E}((\hat{\mu}_k - \mu)^3 | \mathcal{H}_k) > 0$$

$\implies$ $\text{cov}(\hat{\mu}_k, \hat{\tau}_k^2) > 0$

$\implies$ Get small $\hat{\mu}_k$ with small $\hat{\tau}_k^2$ (large $\omega_k$)

and large $\hat{\mu}$ with small $\omega_k$

Result

We would downweight large $\hat{\mu}_k$ (large $\omega_k$)

and upweight small ones

Bad for failure probabilities

Also

$$\text{var}(\hat{\sigma}_k^2 | \mathcal{H}_k) = \infty \text{ possible.}$$

$\hat{\sigma}_k^2 = 0 \text{ possible}$
Model for steady gain

\[ \tau_k^2 = \tau^2 \times k^{-y}, \quad 0 \leq y \leq 1, \quad 0 < \tau < \infty \]

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**

\[ y = 0 \implies \text{no learning} \]

**Too optimistic case**

\[ y = 1 \implies \text{get } \operatorname{var}(\hat{\mu}) = O(N^{-2}) \]

Not reasonable unless \( f(x)p(x) = q(x; \theta) \) some \( \theta \)

We guess \( \tau_k^2 \propto k^x \)

\[ \hat{\mu} = \hat{\mu}(x) = \frac{\sum_{k=1}^{K} k^x \hat{\mu}_k}{\sum_{k=1}^{K} k^x} \quad 0 < x < 1 \]
Variances

\[ \tau_k^2 = \tau^2 k^{-y} \]

\[ \hat{\mu}(x) = \frac{\sum_{k=1}^{K} k^x \hat{\mu}_k}{\sum_{k=1}^{K} k^x} \]

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

At \( x = \text{optimal unknown} \ y \)

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

Rate

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

We should have used $y$ but we did use $x$

$$\rho_K(x \mid y) \equiv \frac{\text{var}(\hat{\mu}(x))}{\text{var}(\hat{\mu}(y))} = \frac{(\sum_{k=1}^{K} k^2x-y)(\sum_{k=1}^{K} ky)}{(\sum_{k=1}^{K} kx)^2}$$

**Just use** $x = 1/2$

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

O & Zhou (2019)

Unknown optimal rate; mildly suboptimal constant.
Steps in the proof

Lemma 1

\[
\sup_{0 \leq y \leq 1} \rho_K(x \mid y) = \begin{cases} 
\rho_K(x \mid 1), & x \leq 1/2 \\
\rho_K(x \mid 0), & x \geq 1/2.
\end{cases}
\]

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

Also: \( \rho_K \left( \frac{1}{2} \mid 0 \right) = \rho_K \left( \frac{1}{2} \mid 1 \right) \).

Lemma 2

\( \rho_{K+1} \left( \frac{1}{2} \mid 1 \right) > \rho_K \left( \frac{1}{2} \mid 1 \right) \), \( K \geq 1 \)

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

Theorem

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

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:

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

Square root rule has

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

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) \( \tau_k^2 \geq \eta > 0 \) for \( k = 1, \ldots, K \)

2) \( \tau_{k+1}^2 \leq \tau_k^2 \)

3) And maybe diminishing returns
   
   (a) \( \tau_{k+2}^2 / \tau_{k+1}^2 \geq \tau_{k+1}^2 / \tau_k^2 \), or
   
   (b) \( \tau_{k+1}^2 - \tau_{k+2}^2 \leq \tau_{k+1}^2 - \tau_k^2 \)

O & Zhou (2019) have some more examples.

Convex minimax

Pick \( \omega_k \) in simplex to

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

Choosing \( \mathcal{T} = \{ (\tau_1^2, \ldots, \tau_K^2) \} \) 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