The Square Root Rule for Adaptive Importance Sampling

ART B. OWEN, Stanford University
YI ZHOU

In adaptive importance sampling and other contexts, we have $K > 1$ unbiased and uncorrelated estimates $\hat{\mu}_k$ of a common quantity μ. The optimal unbiased linear combination weights them inversely to their variances, but those weights are unknown and hard to estimate. A simple deterministic square root rule based on a working model that $\mathrm{Var}(\hat{\mu}_k) \propto k^{-1/2}$ gives an unbiased estimate of μ that is nearly optimal under a wide range of alternative variance patterns. We show that if $\mathrm{Var}(\hat{\mu}_k) \propto k^{-y}$ for an unknown rate parameter $y \in [0, 1]$, then the square root rule yields the optimal variance rate with a constant that is too large by at most $9/8$ for any $0 \leqslant y \leqslant 1$ and any number K of estimates. Numerical work shows that rule is similarly robust to some other patterns with mildly decreasing variance as k increases.

CCS Concepts: • **Computing methodologies** → **Rare-event simulation**;

Additional Key Words and Phrases: Graphics, particle transport, rare events

ACM Reference format:
Art B. Owen and Yi Zhou. 2020. The Square Root Rule for Adaptive Importance Sampling. *ACM Trans. Model. Comput. Simul.* 30, 2, Article 13 (March 2020), 12 pages.
https://doi.org/10.1145/3350426

1 INTRODUCTION

A useful rule for weighting uncorrelated estimates appears in an unpublished technical report [22]. This article states and proves that result and discusses some mild generalizations. The motivating context for the problem is adaptive importance sampling.

Importance sampling is a fundamental Monte Carlo method used in finance [9], reliability [1, 21], coding theory [7], particle transport [12, 13], computer graphics [23, 25], queuing [2], and sequential analysis [14], among other areas.

The prototypical problem is to estimate a constant $\mu = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ for a density function p and integrand f. Although we illustrate the problem with continuous distributions, discrete distributions are handled similarly. In importance sampling, we might estimate μ by

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

This work was supported by the US NSF under grants DMS-1521145 and IIS-1837931. We thank two referees and an associate editor for some extremely valuable comments.
Authors' addresses: A. B. Owen, Department of Statistics, Sequoia Hall, Stanford, CA 94305, USA; email: owen@stanford.edu; Y. Zhou; email: yi_zhou@gmail.com.

1049-3301/2020/03-ART13 $15.00
https://doi.org/10.1145/3350426

The density function q must satisfy $q(\boldsymbol{x};\theta) > 0$ whenever $f(\boldsymbol{x})p(\boldsymbol{x}) \neq 0$, in order to ensure that $\mathbb{E}(\hat{\mu}) = \mu$. In this setting, we must also choose a parameter θ.

A common alternative method is self-normalized importance sampling with

$$\tilde{\mu} = \sum_{i=1}^{n} \frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i;\theta)} \Bigg/ \sum_{i=1}^{n} \frac{p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i;\theta)}, \quad \text{for } \boldsymbol{x}_i \overset{\text{iid}}{\sim} q(\cdot;\theta). \tag{2}$$

This method is more restrictive in requiring $q(\boldsymbol{x};\theta) > 0$ whenever $p(\boldsymbol{x}) > 0$ but less restrictive in that it allows p or q or both to be unnormalized, i.e., known only up to a constant of proportionality. We focus primarily on $\hat{\mu}$, with a few remarks about $\tilde{\mu}$.

After choosing θ and getting $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, we might well decide that some other value of θ would have been better. For instance, we might think that some other θ would yield an estimate of μ with lower variance. In adaptive importance sampling we estimate μ by $\hat{\mu}_k$ at iteration $k \geqslant 1$ using some parameter θ_k. Then, we may use all the sample values from the first k iterations to select θ_{k+1} for the next round.

In an adaptive method with one pilot estimate and one final estimate, $K = 2$ and n is usually large. In other settings, each $\hat{\mu}_k$ could be based on just one evaluation of the integrand (e.g., Ref. [24]), and then K would typically be very large. In intermediate settings, we might have something like $K = 10$ estimates based on perhaps thousands of evaluations each. For instance, the cross-entropy method [6] might be used this way. The total sample size used is $N = nK$. We remark briefly on the possibility that the value of n varies with k in Section 7.

To specify an adaptive importance sampling method, one must come up with a family $q(\cdot;\theta)$ of distributions, a starting value θ_1, a choice for n, an algorithm for computing θ_{k+1} from the data of the prior k rounds, and a termination rule. This article addresses a different issue: after all those choices have been made, how should one pool estimates $\hat{\mu}_1, \ldots, \hat{\mu}_K$ to get the combined estimate $\hat{\mu}$? We have two main points to make:

(1) a natural approach using inverse sample variances to weight $\hat{\mu}_k$ is risky, and
(2) a simple deterministic weighting scheme is nearly optimal over a broad range of reasonable conditions.

We will make the first point with some qualitative arguments and citations.

Most of the article is devoted to the second point showing that a simple deterministic rule is nearly optimal over a set of assumptions ranging from what we consider unduly pessimistic to what we consider unduly optimistic. To that end, we adopt a working model in which $\mathrm{Var}(\hat{\mu}_k) \propto k^{-1/2}$. Using this working model, we get an estimate $\hat{\mu}$ satisfying $\mathbb{E}(\hat{\mu}) = \mu$ along with an unbiased estimate of $\mathrm{Var}(\hat{\mu})$. The working model will ordinarily be incorrect, and the consequence of that incorrectness is a larger value of $\mathrm{Var}(\hat{\mu})$ than would have been attained using the unknown true variances. We give conditions under which the resulting inefficiency is mild enough to be negligible. That lets one avoid a potentially severe challenge of estimating the step-by-step variances and then analyzing a complicated rule derived from those estimates.

This article is organized as follows. Section 2 gives reasons to study settings where $\mathrm{Var}(\hat{\mu}_k)$ will decay steadily but not dramatically, and with diminishing returns as k increases. Section 3 presents our notation and explains why we can assume that the estimates $\hat{\mu}_k$ are unbiased and uncorrelated. Section 4 presents our linear combination estimator, discusses why it is safer to use a prespecified linear combination of $\hat{\mu}_k$ than to use estimated variances, and gives an expression for the relative inefficiency that stems from using a weighting derived from a potentially incorrect power law. Section 5 states and proves our main result using two lemmas. It is that the square root rule never raises the variance by more than 9/8 times the optimal variance for any setting

where $\mathrm{Var}(\hat{\mu}_k) \propto k^{-y}$ for any $0 \leqslant y \leqslant 1$ and any number $K \geqslant 1$ of steps. It is accurate to within that constant factor even though the unknown optimal $\mathrm{Var}(\hat{\mu})$ converges at rates ranging from $O(N^{-1})$ to $O(N^{-2})$. Section 6 considers some alternative ways to model steady improvement with diminishing returns. We find numerically that weighting proportionally to $k^{1/2}$ loses little accuracy under very general circumstances, although some variance patterns give worse than a 9/8 variance ratio. Section 7 includes some remarks on how to weight the $\hat{\mu}_k$ in settings where even better convergence rates, including exponential convergence, apply. Section 8 has brief conclusions.

2 PLAUSIBLE CONVERGENCE RATES

The key assumption we need to make is about how quickly the statistical efficiency of the individual estimates $\hat{\mu}_k$ increases with k. Very commonly $f(\boldsymbol{x}) \geqslant 0$ and then $f(\boldsymbol{x})p(\boldsymbol{x})/\mu$ is a probability density function. If we could sample from that density, then we would have a zero variance estimate $\hat{\mu}$. In special circumstances, $f(\boldsymbol{x})p(\boldsymbol{x})/\mu = q(\boldsymbol{x};\theta)$ for some parameter vector $\theta \in \Theta \subset \mathbb{R}^r$ in an r-dimensional family of densities. It may then be possible to get good estimates of θ_k that improve rapidly as k increases, with the effect that $\mathrm{Var}(\hat{\mu}_k)$ decreases exponentially fast to 0 as k increases. This situation is very well studied in the literature. See Refs. [12], [13], and [18]. We consider it to be especially favorable and not representative of importance sampling in general.

In other settings, $f(\boldsymbol{x})p(\boldsymbol{x})/\mu$ does not take the form $q(\boldsymbol{x};\theta)$ for any $\theta \in \Theta$, and then $\hat{\mu}_k$ cannot have zero variance for finite n. It remains possible that θ_k approaches an optimal vector and then $\mathrm{Var}(\hat{\theta}_k)$ can improve toward a positive lower bound but no further. Exponential convergence is ruled out *a fortiori*.

The self-normalized estimator $\tilde{\mu}_k$ cannot approach zero variance for fixed n outside of trivial settings where $f(\boldsymbol{x})$ is constant under $\boldsymbol{x} \sim p$. It is a ratio estimator and the asymptotically optimal q is proportional to $|f(\boldsymbol{x}) - \mu|p(\boldsymbol{x})$. Any self-normalized estimator satisfies

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

See the Appendix of Ref. [21], which uses the optimal self-normalized importance sampler found by Ref. [11], also given in Ref. [10]. When $f(\boldsymbol{x})$ is the indicator of an event with probability $\epsilon \in (0,1)$ under $\boldsymbol{x} \sim p$, then the asymptotically optimal self-normalized sampler has $\Pr(f(\boldsymbol{x}) = 1; \boldsymbol{x} \sim q) = 1/2$, and any self-normalized importance sampler has

$$\lim_{n\to\infty} n\mathrm{Var}(\tilde{\mu}) \geqslant 4\epsilon^2(1-\epsilon)^2.$$

As a consequence, any nontrivial instance of the self-normalized importance sampler must also have a nonzero floor on its mean squared error for fixed n. This rules out exponential convergence for self-normalized importance sampling in non-trivial problems.

In a third possibility, $q(\boldsymbol{x};\theta)$ is an infinite dimensional family of densities in which $p(\boldsymbol{x})f(\boldsymbol{x})$ can be approached to any desired level of accuracy by some choice of θ. Following up on a strategy of Ref. [26], the nonparametric importance sampler of Zhang [28] has $K = 2$ steps. The second step is an importance sample from a kernel density estimate of $p(\boldsymbol{x})f(\boldsymbol{x})/\mu$ obtained using the first step as a pilot sample. He gives conditions under which the mean squared error at the second stage is $O(N^{-(d+8)/(d+4)})$ for $\boldsymbol{x} \in \mathbb{R}^d$. A multistep adaptive algorithm attains the same rate in N with a different constant that is not necessarily better than using $K = 2$.

We consider exponential convergence to be a rare and special circumstance, and we are most interested in problems where exponential convergence is not possible. We assume that $\mathrm{Var}(\hat{\mu}_k)$ reduces steadily with k but not spectacularly so, and most of our analysis assumes diminishing returns as k keeps increasing. Our main model for steady variance reduction is $\mathrm{Var}(\hat{\mu}_k) \propto k^{-y}$ for an unknown value of y. While the actual variance won't follow this pattern exactly, we expect that

this pattern is qualitatively similar to realistic progress, and it is amenable to sharp analysis. Our combined estimate is a linear combination of $\hat{\mu}_k$. The resulting variance is a linear combination of $\mathrm{Var}(\hat{\mu}_k)$, and so mild departures from a k^{-y} pattern will have little effect on the conclusion.

We consider the case with $y = 0$ to be unduly pessimistic. It describes a setting where adaptation brings no benefit whatsoever. We consider $y = 1$ to be unduly optimistic. While it is not as favorable as the exponential case described above, we show below that it yields an asymptotic mean squared error of $O(N^{-2})$. That is better than the nonparametric method can attain in even one-dimensional problems, and it is much better than we can attain when $p(\boldsymbol{x})f(\boldsymbol{x})/\mu$ cannot be approximated arbitrarily well with some $q(\boldsymbol{x};\theta)$.

When $f(\boldsymbol{x})p(\boldsymbol{x})/\mu$ cannot be arbitrarily well approximated by $q(\boldsymbol{x};\theta)$ for any θ, then it is not actually possible to have $\mathrm{Var}(\hat{\mu}_k) \propto k^{-y}$ for any $y > 0$ and all $k \geqslant 1$. This exponential decay remains a reasonable working model for initial transients over $1 \leqslant k \leqslant K$ for some finite K.

Our proposal is to weight the iterates $\hat{\mu}_k$ as if $y = 1/2$. The result is

$$\hat{\mu} = \sum_{k=1}^{K} \omega_k \hat{\mu}_k \quad \text{for} \quad \omega_k = \frac{\sqrt{k}}{\sum_{i=1}^{K} \sqrt{i}}, \tag{3}$$

a square root rule.

3 NOTATION

Step k of our adaptive importance sampler generates data X_k. This includes all of the sample points $\boldsymbol{x}_{ki}$ along with the function evaluations $f(\boldsymbol{x}_{ki})$, $p(\boldsymbol{x}_{ki})$, and $q(\boldsymbol{x}_{ki};\theta_k)$ at that step. It can also include $q(\boldsymbol{x}_{ki};\theta')$ for values $\theta' \neq \theta_k$. We let $Z_k = (X_1, \ldots, X_{k-1})$ denote the data from all steps prior to the kth, with Z_1 being empty. We assume that our estimates $\hat{\mu}_k$ satisfy

$$\mathbb{E}(\hat{\mu}_k \,|\, Z_k) = \mu, \quad \text{and} \quad \mathrm{Var}(\hat{\mu}_k \,|\, Z_k) = \sigma_k^2 < \infty, \quad k = 1, \ldots, K. \tag{4}$$

This is a reasonable assumption for the estimates in Equation (1), though we are ruling out extreme settings with $\sigma_k^2 = \infty$. The self-normalized estimate $\tilde{\mu}$ of Equation (2) does not satisfy Equation (4) because, as a ratio estimator, it incurs a bias of size $O(1/n)$. For large n, the bias might be small enough that Equation (4) becomes a good approximation for the self-normalized case also.

For the regular importance sampler in Equation (1),

$$\mathrm{Var}(\hat{\mu}_k) = \mathbb{E}(\mathrm{Var}(\hat{\mu}_k \,|\, Z_k)) + \mathrm{Var}(\mathbb{E}(\hat{\mu}_k \,|\, Z_k)) = \mathbb{E}(\sigma_k^2) \equiv \tau_k^2. \tag{5}$$

There is ordinarily an unbiased estimate of σ_k^2, such as

$$\hat{\sigma}_k^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i;\theta_k)} - \hat{\mu}_k \right)^2.$$

Under some moment conditions, $\mathbb{E}(\hat{\sigma}_k^2 \,|\, Z_k) = \sigma_k^2$ and then

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

That is, $\hat{\sigma}_k^2$ is simultaneously an unbiased estimate of both the random variable $\mathrm{Var}(\hat{\mu}_k \,|\, Z_k)$ and the constant $\mathrm{Var}(\hat{\mu}_k)$.

In Ref. [22], $\hat{\mu}_k$ was an importance sampled estimate of an integral over the unit cube, sampling from a mixture of products of beta distributions. Then, θ_k included the mixture components and beta distribution parameters used at step k. That article considers a synthetic integrand formed by a mixture of Gaussians and an integrand with positive and negative singularities from particle physics.

The estimates $\hat{\mu}_k$ are, in general, dependent because data sampled at step k influences the choice of θ_ℓ for $\ell > k$, and thereby affects $\hat{\mu}_\ell$. These estimates are uncorrelated because

$$\begin{aligned}\mathrm{Cov}(\hat{\mu}_k, \hat{\mu}_\ell) &= \mathbb{E}(\mathbb{E}((\hat{\mu}_k - \mu)(\hat{\mu}_\ell - \mu) \,|\, Z_\ell)) \\ &= \mathbb{E}((\hat{\mu}_k - \mu)\mathbb{E}(\hat{\mu}_\ell - \mu \,|\, Z_\ell)) = 0.\end{aligned}$$

The underlying idea in the above expressions is that $\sum_{k=1}^{L} \omega_k(\hat{\mu}_k - \mu)$ is a martingale in L [27], for any fixed set of weights ω_k. We will not make formal use of martingale arguments.

4 LINEAR COMBINATIONS

A simple strategy is to estimate μ by

$$\hat{\mu} = \sum_{k=1}^{K} \omega_k \hat{\mu}_k, \quad \text{where} \quad \sum_{k=1}^{K} \omega_k = 1.$$

For deterministic ω_k, we have $\mathbb{E}(\hat{\mu}) = \mu$ with $\mathrm{Var}(\hat{\mu}) = \sum_{k=1}^{K} \omega_k^2 \tau_k^2$ and then $\widehat{\mathrm{Var}}(\hat{\mu}) = \sum_{k=1}^{K} \omega_k^2 \hat{\sigma}_k^2$ is an unbiased estimate of $\mathrm{Var}(\hat{\mu})$, even if $\mathrm{Var}(\hat{\mu}_k)$ is not proportional to ω_k^{-1}.

To minimize $\mathrm{Var}(\hat{\mu})$, we should take $\omega_k \propto \mathrm{Var}(\hat{\mu}_k)^{-1}$. The problem is that we do not know $\mathrm{Var}(\hat{\mu}_k)$. It is not advisable to plug in the unbiased estimates $\hat{\sigma}_k^2$ and take $\omega_k \propto \hat{\sigma}_k^{-2}$, for several reasons that we outline next.

The importance sampling context is often one where $f(\boldsymbol{x})$ is the indicator of a very rare event. For background on importance sampling for rare events, see Ref. [17]. Another common setting is for $f(\boldsymbol{x})$ to be a nonnegative random variable with a very heavy right tail, perhaps even an integrable singularity. Both of these cases describe a setting where $f(\boldsymbol{x})$ has extremely large positive skewness under $\boldsymbol{x} \sim p$. We then expect to get $f(\boldsymbol{x})p(\boldsymbol{x})/q(\boldsymbol{x})$ with large positive skewness under $\boldsymbol{x} \sim q$. Evans and Swartz [8] describe this as low effective sample size for the integrand f. The more general low effective sample size problem shows up as extreme positive skewness in $p(\boldsymbol{x})/q(\boldsymbol{x})$ under $\boldsymbol{x} \sim q$.

In cases of positive skewness, we get a positive correlation between the sample values of $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ [19]. Weighting points proportionally to $\hat{\sigma}_k^{-2}$ then means that the largest $\hat{\mu}_k$ tend to be randomly downweighted while the smallest ones tend to be randomly upweighted. The result is a negative bias in $\hat{\mu}$, which is especially undesirable when μ is the probability of a rare adverse event. For a rare event that fails to happen even once at step k, we could get $\hat{\mu}_k = \hat{\sigma}_k^2 = 0$, and then some complicated intervention would be required to make the formulas work. It is possible that some settings will have negative skewness for fp/q and a corresponding upward bias using estimated variances, though that is less to be expected in importance sampling. This could happen for problems with negative integrable singularities or problems with both positive and negative integrable singularities or when estimating the probability of an event whose complement is rare.

In addition to the bias issue, there is a possibly more fundamental reason not to use $\hat{\sigma}_k^2$ in weights. Importance sampling is often used in settings where simply estimating μ is quite hard. Estimating σ_k^2 or τ_k^2 well could then be an order of magnitude harder. They only have finite variances under a finite fourth moment condition for fp/q. Using estimates of those quantities then amounts to using an extremely noisy weighting relying on a subsidiary problem (variance estimation) that is harder than the motivating problem. See Ref. [15] for an analysis of how hard it is to estimate variance in the importance sampling context. This issue is especially severe in cases with large K and small n. Chatterjee and Diaconis [4] also remark on the difficulty of using variances in importance sampling.

The adaptive multiple importance sampling (AMIS) algorithm of Ref. [5] uses a sample determined weighted combination of estimates. In that case, the weight applied to $\hat{\mu}_k$ can depend on

observations from future rounds in a more complicated way than just normalizing $\hat{\sigma}_k^{-2}$. The intricate dependence pattern between weights $\hat{\omega}_k$ and estimates $\hat{\mu}_k$ complicates even the task of proving consistency of the estimator. Their setup was for self-normalized importance sampling, but the difficulties with noisy estimates of optimal weights apply there, too. Even without that complexity, taking account of sampling fluctations in both $\hat{\mu}_k$ and $\hat{\omega}_k$ can raise challenging issues that a deterministic rule avoids.

We are interested in cases where adaptation brings steady but not dramatic improvements. As noted above, our model for that is

$$\mathrm{Var}(\hat{\mu}_k) = \tau^2 k^{-y} \tag{6}$$

for $0 \leqslant y \leqslant 1$ and $\tau = \tau_1 \in (0, \infty)$. If we worked with $\tau^2 k^{-x}$, then our estimate would be

$$\hat{\mu} = \hat{\mu}(x) = \sum_{i=1}^{K} i^x \hat{\mu}_i \Big/ \sum_{i=1}^{K} i^x,$$

including our Proposal (3) via $x = 1/2$. The best choice is $\hat{\mu}(y)$, but y is unknown. Our variance using x is

$$\mathrm{Var}(\hat{\mu}(x)) = \tau^2 \sum_{i=1}^{K} i^{2x-y} \Big/ \left(\sum_{i=1}^{K} i^x \right)^2,$$

and we measure the inefficiency of our choice by

$$\rho_K(x \,|\, y) \equiv \frac{\mathrm{Var}(\hat{\mu}(x))}{\mathrm{Var}(\hat{\mu}(y))} = \frac{\left(\sum_{i=1}^{K} i^{2x-y}\right)\left(\sum_{i=1}^{K} i^{y}\right)}{\left(\sum_{i=1}^{K} i^{x}\right)^2}. \tag{7}$$

If the variance of $\hat{\mu}_k$ decays as k^{-y} and, knowing that, we use $x = y$, then $\mathrm{Var}(\hat{\mu}) = O(K^{-y-1}) = O(N^{-y-1})$. The value $y = 0$ is pessimistic as it corresponds to a setting where adaptation brings no benefits. For an optimistic value $y = 1$, the variance decays at the rate $O(N^{-2})$, which we consider unreasonably optimistic. It is better than the rate that Ref. [28] gets for importance sampling by one-dimensional kernel density estimates and even slightly better than the rate that holds for randomly shifted lattice rules applied to functions of bounded variation in the sense of Hardy and Krause. See Ref. [16] for background on randomized lattice rules.

5 SQUARE ROOT RULE

In Ref. [22], we proposed to split the difference between the optimistic estimate $\hat{\mu}(1)$ and the pessimistic one $\hat{\mu}(0)$ by using $\hat{\mu}(1/2)$. The result is an unbiased estimate that attains the same convergence rate as the unknown best estimate with only a modest penalty in the constant factor.

Theorem 1. *For $\rho_K(x \,|\, y)$ given by Equation (7),*

$$\sup_{1 \leqslant K < \infty} \sup_{0 \leqslant y \leqslant 1} \rho_K\left(\frac{1}{2} \,\Big|\, y\right) \leqslant \frac{9}{8}. \tag{8}$$

If $x \neq 1/2$ and $K \geqslant 2$, then

$$\sup_{0 \leqslant y \leqslant 1} \rho_K(x \,|\, y) > \sup_{0 \leqslant y \leqslant 1} \rho_K\left(\frac{1}{2} \,\Big|\, y\right). \tag{9}$$

Equation (8) shows that the square root rule has at most 12.5% more variance than the unknown best rule. Equation (9) shows that the choice $x = 1/2$ is the unique minimax optimal one, when $K > 1$. When $K = 1$, then $\hat{\mu}(x)$ does not depend on x, and, in that case, $\hat{\mu}(x) = \hat{\mu}(y)$. Before proving the theorem, we establish two lemmas.

Lemma 1. *For $\rho_K(x\,|\,y)$ given by Equation (7), with $K \geqslant 1$ and $0 \leqslant x \leqslant 1$,*

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

Proof. The result holds trivially for $K = 1$ because $\rho_1(x\,|\,y) = 1$. Now suppose that $K \geqslant 2$. Then,

$$\frac{\partial^2}{\partial y^2}\rho_K(x\,|\,y) = C^{-2} \times \sum_{i=1}^{K}\sum_{j=1}^{K} i^{2x-y} j^{y} (\log(i) - \log(j))^2$$

for $C = \sum_{i=1}^{K} i^x$. Thus, $\rho_K(x\,|\,y)$ is strictly convex in y for any x when $K \geqslant 2$, and so $\sup_{0\leqslant y\leqslant 1} \rho_K(x\,|\,y) \in \{\rho_K(x\,|\,1), \rho_K(x\,|\,0)\}$.

By symmetry, $\rho_K(x\,|\,2x - y) = \rho_K(x\,|\,y)$. So if $x \geqslant 1/2$, then $\rho_K(x\,|\,0) = \rho_K(x\,|\,2x) \geqslant \rho_K(x, 1)$. The last step follows because $2x \geqslant 1$ and $\rho_K(x\,|\,y)$ is a convex function of y with its minimum at $y = x \leqslant 1$. This establishes the result for $x \geqslant 1/2$ and a similar argument holds for $x \leqslant 1/2$. For $x = 1/2$, $\rho_K(x\,|\,0) = \rho_K(x\,|\,1)$. □

Proposition 1 has two integral bounds that we will use below.

Proposition 1. *If $0 \leqslant x \leqslant 1$, and $K \geqslant 1$ is an integer, then*

$$\sum_{i=1}^{K} i^x \geqslant \int_{1/2}^{K+1/2} v^x \,\mathrm{d}v = \frac{(K+1/2)^{x+1} - (1/2)^{x+1}}{x+1}, \quad \textit{and} \tag{10}$$

$$\sum_{i=1}^{K} i^x \leqslant \int_{1}^{K+1} v^x \,\mathrm{d}v = \frac{(K+1)^{x+1} - 1}{x+1}. \tag{11}$$

Equation (10) is strict when $0 < x < 1$. Equation (11) is strict when $0 < x \leqslant 1$.

Proof. Equation (10) follows from the concavity of v^x. That concavity is strict for $0 < x < 1$. Equation (11) follows because $v^x \geqslant i^x$ for $i \leqslant v \leqslant i + 1$, which holds strictly for $x > 0$. □

Equation (10) is much sharper than the bound one gets by integrating v^x over $0 \leqslant v \leqslant K$.

Lemma 2. *For $\rho_K(x\,|\,y)$ given by Equation (7), $\rho_{K+1}(1/2\,|\,1) > \rho_K(1/2\,|\,1)$ holds for any integer $K \geqslant 1$.*

Proof. Let $S_K = \sum_{i=1}^{K} i^{1/2}$. Then, using Equation (10) from Proposition 1,

$$\begin{aligned}
\frac{\rho_{K+1}(1/2\,|\,1)}{\rho_K(1/2\,|\,1)} &= \frac{(K+1)(K+2)S_K^2}{K^2(S_K + \sqrt{K+1})^2} = \frac{(K+1)(K+2)}{K^2\left(1 + \frac{\sqrt{K+1}}{S_K}\right)^2} \\
&> \frac{(K+1)(K+2)}{K^2} \Bigg/ \left(1 + \frac{\sqrt{K+1}}{((K+1/2)^{3/2} - 2^{-3/2})/(3/2)}\right)^2 \\
&= (K+1)(K+2) \Bigg/ \left(K + \frac{3}{2}\frac{K}{f(K) - 2^{-3/2}/\sqrt{K+1}}\right)^2,
\end{aligned}$$

for

$$f(K) = (K+1/2)^{3/2}/\sqrt{K+1}.$$

Let $g(K) = f(K)/(K+1/4)$. We easily find that $\lim_{K\to\infty} g(K) = 1$. Also, g is monotone decreasing in K because the derivative of $\log(g(K))$ with respect to K is $-(3/16)[(K+1/2)(K+1/4)(K+1)]^{-1}$. Therefore, if $K > 7$,

$$\begin{aligned}\frac{\rho_{K+1}(1/2\,|\,1)}{\rho_K(1/2\,|\,1)} &> (K+1)(K+2)\left(K+\frac{3}{2}\frac{K}{g(K)(K+1/4)-2^{-3/2}/\sqrt{8}}\right)^{-2}\\ &> (K+1)(K+2)\left(K+\frac{3}{2}\frac{K}{K+1/4-2^{-3/2}/\sqrt{8}}\right)^{-2}\\ &= \frac{(K+1)(K+2)}{\left(K+\frac{3}{2}\frac{K}{K+1/8}\right)^2}. \end{aligned} \tag{12}$$

The numerator in Equation (12) is K^2+3K+2, while the denominator is

$$K^2 + 3K\frac{K}{K+1/8} + \frac{9}{4}\frac{K^2}{(K+1/8)^2}.$$

The numerator is larger than the denominator, establishing the theorem for $K > 7$. For $1 \leqslant K \leqslant 7$, we compute directly that

$$\frac{\rho_{K+1}(1/2\,|\,1)}{\rho_K(1/2\,|\,1)} > 1.0038. \qquad \square$$

Proof of Theorem 1. Applying Lemmas 1 and 2,

$$\sup_{1\leqslant K<\infty}\sup_{0\leqslant y\leqslant 1}\rho_K\left(\frac{1}{2}\,\middle|\,y\right) = \lim_{K\to\infty}\rho_K\left(\frac{1}{2}\,\middle|\,1\right) = \lim_{K\to\infty}\frac{K^2(K+1)/2}{\left(\sum_{i=1}^K i^{1/2}\right)^2}.$$

Using the integral Bounds (10) and (11) from Proposition 1, we can show that the above limit is $9/8$, establishing Equation (8). Next, if $x < 1/2$, then

$$\begin{aligned}\sup_{0\leqslant y\leqslant 1}\rho_K(x\,|\,y) = \rho_K(x\,|\,1) &= \frac{\left(\sum_{i=1}^K i^{2x-1}\right)\left(\sum_{i=1}^K i\right)}{\left(\sum_{i=1}^K i^x\right)^2}\\ &= \frac{K(K+1)}{2}\frac{\left(\sum_{i=1}^K i^{2x-1}\right)}{\left(\sum_{i=1}^K i^x\right)^2}.\end{aligned}$$

Ignoring the factor $K(K+1)$ and taking the derivative with respect to x yields

$$\frac{\sum_{i=1}^K\sum_{j=1}^K i^x j^{2x-1}(\log(j)-\log(i))}{\left(\sum_{i=1}^K i^x\right)^3}. \tag{13}$$

The numerator in Equation (13) is of the form $\sum_{\ell=1}^K \eta_\ell \log(\ell)$ where $\sum_{\ell=1}^K \eta_\ell = 0$. That quantity η_ℓ is $\sum_{i=1}^K (i^x\ell^{2x-1} - \ell^x i^{2x-1})$, which is a decreasing function of ℓ because $x < 1/2$. Next, because $\log(\ell)$ is non-negative and increasing in ℓ, we find that the numerator in Equation (13) is negative and, hence, so is the whole expression. Therefore, $\rho_K(x\,|\,1)$ is a decreasing function of x making $\rho_K(x\,|\,1) > \rho_K(1/2\,|\,1)$ for $x < 1/2$. This establishes Equation (9) for $x < 1/2$, and the case of $x > 1/2$ is similar and slightly easier. $\square$

6 ROBUSTNESS

It is not reasonable to suppose that even the more general working model $\mathrm{Var}(\hat{\mu}_k) \propto k^{-y}$ would hold exactly. Using the square root rule

$$\mathrm{Var}(\hat{\mu}) = \sum_{k=1}^{K} k\mathrm{Var}(\hat{\mu}_k) \Big/ \left(\sum_{k=1}^{K} k^{1/2} \right)^2$$

is a continuous function of $\mathrm{Var}(\hat{\mu}_k)$. Then, small perturbations from a power law make a small difference. Also a positive function of k that decreases mildly and steadily and convexly as k increases can be expected to be highly correlated with k^{-y} for some y. The inefficiency of the square root rule in this more general setting is

$$\rho = \sum_{k=1}^{K} k\mathrm{Var}(\hat{\mu}_k) \sum_{k=1}^{K} \mathrm{Var}(\hat{\mu}_k)^{-2} \Big/ \left(\sum_{k=1}^{K} k^{1/2} \right)^2. \tag{14}$$

One plausible alternative to the power law is that $\mathrm{Var}(\hat{\mu}_k)$ might make steady progress and then plateau. An example of that would be $\mathrm{Var}(\hat{\mu}_k) = \min(k, 6)^{-1}$ for $k = 1, \dots, 10$. We can compute ρ of Equation (14) directly in this case and find that using $\omega_k \propto k^{1/2}$ would increase variance by a factor of just under 1.04 compared to using the optimal weights. If we suppose that there are $k_1 \in \{1, \dots, 100\}$ stages with variance decreasing proportionally to k^{-1} followed by a plateau of $k_2 \in \{1, \dots, 100\}$ steps with variance proportional to $(k_1 + 1)^{-1}$, then the square root rule never raises variance by more than 1.121 over the optimal combination.

A more general form of mild improvement has $\sigma_k^2 \geqslant 1/k$ to rule out unreasonably good performance, $\sigma_{k+1}^2 \leqslant \sigma_k^2$ to model improvement, and $\sigma_k^2 - \sigma_{k+1}^2 \geqslant \sigma_{k+1}^2 - \sigma_{k+2}^2$ to model diminishing returns. Taking $\sigma_1^2 = 1$ and a fixed K, these conditions yield a convex set of vectors $(\sigma_1^2, \dots, \sigma_K^2)$. It would be interesting to know the largest inefficiency ρ over that set of possibilities, but this ρ is not a concave function of the σ_k^2 values, and finding the exact worst case is outside the scope of this article.

We can sample the space of possible values by recursively taking $\sigma_1^2 = 1$, $\sigma_2^2 \sim U[1/2, \sigma_1^2]$, and for $2 \leqslant i \leqslant K$,

$$\sigma_k^2 \sim U\Big[\min(1/k, 2\sigma_{k-1}^2 - \sigma_{k-2}^2),\ \sigma_{k-1}^2\Big].$$

In 10^6 simulations with $K = 5$, the inefficiency was worse than 9/8 only 18 times, and the worst value was about 1.134. In 10^6 simulations with $K = 10$, the inefficiency was worse than 9/8 1,904 times, and the worst value was about 1.297. In 10^6 simulations with $K = 20$, the inefficiency was worse than 9/8 101 times, and the worst value was about 1.263. There is not much to lose in weighting proportionally to $k^{1/2}$ in this diminishing returns setup.

If the progress is not steady and convex, then the square root rule can be more inefficient than described above. If the variance remains flat for the first k_1 steps and then decreases sharply, the square root rule is less efficient. For instance, if the first 3 iterations have unit variance and the next 10 have variance 0.01, then the square root rule raises variance by about 6.37 fold over the optimal weights. In Ref. [22], we advocated putting zero weight on the first few iterations if such an initial transient is suspected.

7 GENERALIZATION

We can generalize some aspects of the square root rule to other power laws. Suppose that the true rate parameter y is known to satisfy $L \leqslant y \leqslant U$ for $0 \leqslant L \leqslant U < \infty$. We can then work with

$\omega_k \propto k^x$ for $x = M \equiv (L+U)/2$. First, we recall the definition of ρ_K from Equation (7):

$$\rho_K(x\,|\,y) = \frac{\left(\sum_{i=1}^K i^{2x-y}\right)\left(\sum_{i=1}^K i^y\right)}{\left(\sum_{i=1}^K i^x\right)^2}.$$

Then, $2M - U = L$, and

$$\rho_K(M\,|\,U) = \frac{\left(\sum_{i=1}^K i^L\right)\left(\sum_{i=1}^K i^U\right)}{\left(\sum_{i=1}^K i^M\right)^2}. \tag{15}$$

Lemma 3. *For $\rho_K(x\,|\,y)$ given by Equation (7), with $K \geqslant 1$ and $0 \leqslant L \leqslant x \leqslant U < \infty$,*

$$\sup_{L\leqslant y\leqslant U} \rho_K(x\,|\,y) = \begin{cases} \rho_K(x\,|\,U), & x \leqslant M \\ \rho_K(x\,|\,L), & x \geqslant M, \end{cases}$$

for $M = (L+U)/2$.

Proof. The proof is the same as for Lemma 1 because $\rho_K(x\,|\,y)$ is convex in y for any $K \geqslant 1$ and any x and $\rho_K(x\,|\,2x-y) = \rho_K(x\,|\,y)$. □

The inequalities in Lemma 2 are rather delicate, and we have not extended them to the more general setting. Equation (15) is easy to evaluate for integer values of L, M, and U. For more general values, some sharper tools than the integral bounds in this article are given by Ref. [3] who make a detailed study of sums of powers of the first K natural numbers. We do see, numerically, that $\rho_K(M\,|\,U)$ is nondecreasing with K in every instance we have inspected. Using Proposition 1, we can easily find the asymptotic inefficiency

$$\lim_{K\to\infty} \rho_K(M\,|\,U) = \lim_{K\to\infty} \frac{K^L}{L+1}\frac{K^U}{U+1} \Big/ \left(\frac{K^M}{M+1}\right)^2 = \frac{(M+1)^2}{(L+1)(U+1)}.$$

In cases with $L = 0$ and, hence, $M = U/2$, we get $(U/2+1)^2/(U+1)$. For instance, an upper bound at the rate $y = U = 2$ corresponding roughly to asymptotic accuracy of scrambled net integration [20] leads to $x = M = 1$ and an asymptotic inefficiency of at most

$$\frac{(1+1)^2}{(0+1)(2+1)} = \frac{4}{3}.$$

Some references in Section 2 describe problems where the adaptive importance sampling variance converges exponentially to zero. It is reasonable to expect that each estimate $\hat\mu_k$ will require a large number n of observations to fit a complicated integrand, and that K will then be not too large.

Suppose that $\mathrm{Var}(\hat\mu_k) = \tau^2 \exp(-yk)$ for some $y > 0$. Then, the desired combination is

$$\hat\mu(y) = \frac{\sum_{i=1}^K \exp(iy)\hat\mu_i}{\sum_{i=1}^K \exp(iy)}.$$

Not knowing y, we use

$$\hat\mu(x) = \frac{\sum_{i=1}^K \exp(ix)\hat\mu_i}{\sum_{i=1}^K \exp(ix)},$$

for some $x > 0$. The inefficiency of our estimate is now

$$\gamma_K(x\,|\,y) \equiv \frac{\sum_{i=1}^K \exp((2x-y)i)\sum_{i=1}^K \exp(yi)}{\left(\sum_{i=1}^K \exp(xi)\right)^2}$$
$$= \left(\frac{e^{2x-y}(e^{K(2x-y)}-1)}{e^{2x-y}-1}\right)\left(\frac{e^{y}(e^{Ky}-1)}{e^{y}-1}\right)\Big/\left(\frac{e^{x}(e^{Kx}-1)}{e^{x}-1}\right)^2.$$

If $x = y/2$, then the first factor in the numerator is K. It can be disastrously inefficient to use $x \ll y$. Some safety is obtained in the $x \to \infty$ limit where $\hat{\mu} = \mu_K$. In that limit, we consider using the final and presumably best estimate to be a sensible default. Then,

$$\lim_{x\to\infty} \gamma_K(x\,|\,y) = \frac{e^y(1-e^{-Ky})}{e^y-1} \leqslant \frac{e^y}{e^y-1}.$$

We have no direct experience with problems having geometric convergence. For sake of illustration, if the variance is halving at each iteration, then $y = \log(2)$ and then taking $x \to \infty$ is inefficient by at most a factor of 2. Repeated ten-fold variance reductions correspond to $y = \log(10)$ and a limiting inefficiency of at most $10/9 \doteq 1.11$. The greatest inefficiency from using only the final iteration arises in the limit $y \to 0$ where the factor is K. In this setting, the user is not getting a meaningful exponential convergence; even there, the loss factor is at most K and, as remarked above, that K is not likely to be large.

We have assumed that n is the same at every iteration. It is possible to take a varying number n_k of observations at iteration k. Zhang [28] obtains his convergence rates for $K = 2$ assuming that n_1 and n_2 are chosen in an asymptotic regime where $n_1/(n_1+n_2)$ approaches a limit in the interior of the interval $(0, 1)$. Somewhat unequal values of n_1 and n_2 optimized the lead constant.

If $\mathrm{Var}(\hat{\mu}_k) = \sigma_k^2/n_k$ with $\sigma_k^2 \propto k^{-y}$ for $0 \leqslant y \leqslant 1$ and $n_k \propto k^{\alpha}$, then $\mathrm{Var}(\hat{\mu}_k) \propto k^{-y-\alpha}$. We might then be able to use the more general bounds above with $L = \alpha$ and $U = 1 + \alpha$. Conceivably, knowing σ_k^2 would let one tune n_k. However, changes to n_k are likely to change σ_k^2, so that problem is circular and solving that is outside the scope of this article.

8 CONCLUSIONS

Some adaptive importance samplers can have geometric convergence, but many other settings will not converge that quickly. A simple rule weighting the estimate at stage k proportionally to $k^{1/2}$ gives an unbiased and nearly efficient estimate under a wide range of conditions and it also lets the user avoid trying to model a random linear combination of $\hat{\mu}_k$. An interesting extension would be to specify a well-motivated convex set of values for $(1, \sigma_2^2/\sigma_1^2, \ldots, \sigma_K^2/\sigma_1^2)$ and find a set of weights $\omega_k \geqslant 0$ with $\sum_k \omega_k = 1$ that minimizes the worst-case inefficiency

$$\max \sum_{k=1}^K \omega_k^2 \mathrm{Var}(\hat{\mu}_k) \sum_{k=1}^K \mathrm{Var}(\hat{\mu}_k)^{-2},$$

where the maximum is taken over variances with ratios in the aforementioned convex set.

REFERENCES

[1] Siu-Kui Au and James L. Beck. 1999. A new adaptive importance sampling scheme for reliability calculations. *Structural Safety* 21, 2 (1999), 135–158.

[2] José Blanchet, Peter Glynn, and Jingchen C. Liu. 2007. Fluid heuristics, Lyapunov bounds and efficient importance sampling for a heavy-tailed G/G/1 queue. *Queueing Systems* 57, 2–3 (2007), 99–113.

[3] Brian L. Burrows and Richard F. Talbot. 1984. Sums of powers of integers. *The American Mathematical Monthly* 91, 7 (1984), 394–403.

[4] Sourav Chatterjee and Persi Diaconis. 2018. The sample size required in importance sampling. *The Annals of Applied Probability* 28, 2 (2018), 1099–1135.
[5] Jean-Marie Cornuet, Jean-Michel Marin, Antonietta Mira, and Christian P. Robert. 2012. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics* 39, 4 (2012), 798–812.
[6] Pieter-Tjerk De Boer, Dirk P. Kroese, Shie. Mannor, and Reuven Y. Rubinstein. 2005. A tutorial on the cross-entropy method. *Annals of Operations Research* 134, 1 (2005), 19–67.
[7] Victor Elvira and Ignacio Santamaria. 2019. Multiple importance sampling for efficient symbol error rate estimation. *IEEE Signal Processing Letters* 26, 3 (2019), 420–424.
[8] Michael J. Evans and Tim Swartz. 1995. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statist. Sci.* 10, 3 (1995), 254–272.
[9] Paul G. Glasserman. 2004. *Monte Carlo Methods in Financial Engineering*. Springer, New York.
[10] Timothy C. Hesterberg. 1988. *Advances in Importance Sampling*. Ph.D. Dissertation. Stanford University.
[11] Herman Kahn and Andrew Marshall. 1953. Methods of reducing sample size in monte carlo computations. *Journal of the Operations Research Society of America* 1, 5 (1953), 263–278.
[12] Craig Kollman, Keith Baggerly, Dennis Cox, and Rick Picard. 1999. Adaptive importance sampling on discrete Markov chains. *Annals of Applied Probability* 9, 2 (1999), 391–412.
[13] Rong Kong and Jerome Spanier. 2011. Geometric convergence of adaptive monte carlo algorithms for radiative transport problems based on importance sampling methods. *Nuclear Science and Engineering* 168, 3 (2011), 197–225.
[14] Tze L. Lai and David Siegmund. 1977. A nonlinear renewal theory with applications to sequential analysis I. *The Annals of Statistics* 5, 5 (1977), 946–954.
[15] Pierre. L'Ecuyer, José. H. Blanchet, Bruno. Tuffin, and Peter W. Glynn. 2010. Asymptotic robustness of estimators in rare-event simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 20, 1 (2010), 6.
[16] Pierre L'Ecuyer and Christian Lemieux. 2000. Variance reduction via lattice rules. *Management Science* 46, 9 (2000), 1214–1235.
[17] Pierre L'Ecuyer, Michel Mandjes, and Bruno Tuffin. 2009. Importance sampling and rare event simulation. In *Rare Event Simulation Using Monte Carlo Methods*, G. Rubino and B. Tuffin (Eds.). John Wiley & Sons, Chichester, UK, 17–38.
[18] Pierre L'Ecuyer and Bruno Tuffin. 2008. Approximate zero-variance simulation. In *Proceedings of the 2008 Winter Simulation Conference*, S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler (Eds.). IEEE Press, 170–181.
[19] Rupert G. Miller. 1986. *Beyond ANOVA: Basics of Applied Statistics*. Wiley, New York.
[20] Art B. Owen. 1997. Scrambled net variance for integrals of smooth functions. *Annals of Statistics* 25, 4 (1997), 1541–1562.
[21] Art B. Owen, Yury Maximov, and Michael Chertkov. 2019. Importance sampling the union of rare events with an application to power systems analysis. *Electronic Journal of Statistics* 13, 1 (2019), 231–254.
[22] Art B. Owen and Y. Zhou. 1999. *Adaptive Importance Sampling by Mixtures of Products of Beta Distributions*. Technical Report. Stanford University.
[23] Matt Pharr, Wenzel Jakob, and Greg Humphreys. 2016. *Physically Based Rendering: From Theory to Implementation* (third ed.). Morgan Kaufmann, Cambridge, MA.
[24] Ernest K. Ryu and Stephen P. Boyd. 2014. *Adaptive Importance Sampling via Stochastic Convex Programming*. Technical Report. arXiv:1412.4845.
[25] Eric Veach and Leonidas Guibas. 1995. Optimally combining sampling techniques for monte carlo rendering. In *SIGGRAPH'95 Conference Proceedings*. Addison-Wesley, Reading, MA, 419–428.
[26] Mike West. 1993. Approximating posterior distributions by mixtures. *Journal of the Royal Statistical Society: Series B* 55, 2 (1993), 409–422.
[27] David Williams. 1991. *Probability with Martingales*. Cambridge University Press, Cambridge.
[28] Ping Zhang. 1996. Nonparametric importance sampling. *J. Amer. Statist. Assoc.* 91, 435 (1996), 1245–1253.

Received January 2019; revised June 2019; accepted July 2019