ADAPTIVE SIMPLE MONTE CARLO

FRED J. HICKERNELL, LAN JIANG, AND YUEWEI LIU

Abstract. We attempt a probabilistic analysis of simple Monte Carlo, achieving probabilistic error bounds when the kurtosis is controlled. The algorithm uses a sample size that depends adaptively on the estimated variance of the integrand. Thus, the algorithm is nonlinear (depending essentially on the function). The advantage of what is done here over standard error analysis (complexity theory) is that the algorithm does not depend a priori on the scale of the problem (in this case the variance) to determine the number of samples. Our intention, if what is done here is correct, is to try to extend this to the more sophisticated sampling schemes and infinite dimensional problems.

1. Simple Monte Carlo

Suppose one wishes to compute the following integral or mean, μ, of some function $f : \mathbb{R}^d \to \mathbb{R}$, i.e.,

$$\mu = \mu(f) = \int_{\mathbb{R}^d} f(\boldsymbol{x})\rho(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$

where $\rho : \mathbb{R}^d \to [0,\infty)$ is a probability density function. A simple Monte Carlo algorithm to estimate this integral is to generate $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ independent and identically distributed random variables with the probability density function ρ, i.e., $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ i.i.d. $\sim \rho$, evaluate the integrand at these sample points, and then take the sample mean:

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i, \qquad Y_i = f(\boldsymbol{X}_i). \tag{1}$$

This is an unbiased estimate, i.e., $E(\hat{\mu}_n) = \mu$. Thus, the mean square error is the variance of $\hat{\mu}_n$:

$$\mathrm{MSE}(\hat{\mu}_n) = E[(\mu - \hat{\mu}_n)^2] = \mathrm{var}(\hat{\mu}_n) = \frac{\mathrm{var}(Y_1)}{n}. \tag{2}$$

The variance of Y_1 may be written in terms of an integral of the f. Define the $\mathcal{L}_p$ norm of f as follows:

$$\|f\|_p := \left\{\int_{\mathbb{R}^d} |f(\boldsymbol{x})|^p \rho(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right\}^{1/p}$$

Note that if $1 \le q < p$, then by Hölder's inequality,

$$\begin{aligned}\|f\|_q &= \left\{\int_{\mathbb{R}^d} |f(\boldsymbol{x})|^q \rho(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right\}^{1/q} \\ &\le \left\{\int_{\mathbb{R}^d} |f(\boldsymbol{x})|^p f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right\}^{1/p} \left\{\int_{\mathbb{R}^d} 1^{p/(p-q)} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right\}^{(p-q)/(pq)} \\ &= \|f\|_p \|1\|_{pq/(p-q)} = \|f\|_p .\end{aligned} \tag{3}$$

Thus, $\mathcal{L}_q \subseteq \mathcal{L}_p$ for $1 \le q < p$.

Now define the the centered absolute p^{th} moment of f as

$$M_p(f) := \int_{\mathbb{R}^d} |f(\boldsymbol{x}) - \mu|^p \, \rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \qquad p \geq 1 \tag{4}$$

By this definition, the variance of Y_1 is the second moment, i.e.,

$$\sigma^2 := \text{var}(Y_1) = \text{var}(f(\boldsymbol{X}_1)) =: \text{var}(f) = M_2(f) = \int_{\mathbb{R}^d} |f(\boldsymbol{x}) - \mu|^2 \, \rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{5}$$

Moreover, by (3) it follows that low order moments can be bounded above by higher order moments.

$$M_q \leq M_p^{q/p} \quad \text{and} \quad M_p < \infty \implies M_q < \infty \qquad \text{for } 1 \leq q \leq p. \tag{6}$$

Thus, the root mean square error of the simple Monte Carlo quadrature rule, given in (2), may be written in terms of the second moment of f as

$$\text{RMSE}(\hat{\mu}_n) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\text{var}(f)}{n}}. \tag{7}$$

2. Randomized Error Analysis

Given an error tolerance, ε, and a set of functions, $\mathcal{F}$, one may ask how large a sample size is needed to guarantee that this tolerance is achieved. The answer depends on the error criterion. Here are three possibilities defined for our simple Monte Carlo (sMC) algorithm:

$$\text{randomized absolute} \quad \text{err}^{\text{rnd abs}}(n, \mathcal{F}, \text{sMC}) := \sup_{f \in \mathcal{F}} \text{RMSE}(\hat{\mu}_n) = \sqrt{\frac{\sup_{f \in \mathcal{F}} \text{var}(f)}{n}}, \tag{8a}$$

$$\text{randomized relative} \quad \text{err}^{\text{rnd rel}}(n, \mathcal{F}, \text{sMC}) := \sup_{f \in \mathcal{F}} \frac{\text{RMSE}(\hat{\mu}_n)}{\mu} = \frac{1}{\sqrt{n}} \sup_{f \in \mathcal{F}} \frac{\sqrt{\text{var}(f)}}{\mu(f)}, \tag{8b}$$

$$\text{randomized normalized} \quad \text{err}^{\text{rnd nor}}(n, \mathcal{F}, \text{sMC}) := \sup_{f \in \mathcal{F}} \frac{\text{RMSE}(\hat{\mu}_n)}{\sqrt{\text{var}(f)}} = \frac{1}{\sqrt{n}}. \tag{8c}$$

Here the root mean square error is taken over all $\hat{\mu}_n$ with cardinality n. Given an error criterion, err^{x}, the sample size needed to attain the error tolerance, ε. for this class of algorithms with deterministic sample size is defined as

$$\text{cost}^{\text{x}}(\varepsilon, \mathcal{F}, \text{sMC}) = \min\{n \in \mathbb{N}_0 : \text{err}^{\text{x}}(n, \mathcal{F}, \text{sMC}) \leq \varepsilon\},$$

where $\text{x} \in \{\text{rnd abs}, \text{rnd rel}, \text{rnd nor}\}$. For the error criteria in (8) and the simple Monte Carlo algorithm, it is found that

$$\text{cost}^{\text{rnd abs}}(\varepsilon, \mathcal{F}^{\text{abs}}, \text{sMC}) = \frac{\sigma_{\max}^2}{\varepsilon^2}, \qquad \mathcal{F}^{\text{abs}} = \left\{f \in \mathcal{L}_2 : \text{var}(f) \leq \sigma_{\max}^2\right\}, \tag{9a}$$

$$\text{cost}^{\text{rnd rel}}(\varepsilon, \mathcal{F}^{\text{rel}}, \text{sMC}) = \frac{a^2}{\varepsilon^2}, \qquad \mathcal{F}^{\text{rel}} = \left\{f \in \mathcal{L}_2 : \frac{\text{var}(f)}{[\mu(f)]^2} \leq a^2, \ \mu(f) \neq 0\right\}, \tag{9b}$$

$$\text{cost}^{\text{rnd nor}}(\varepsilon, \mathcal{F}^{\text{nor}}, \text{sMC}) = \frac{1}{\varepsilon^2}, \qquad \mathcal{F}^{\text{nor}} = \mathcal{L}_2. \tag{9c}$$

The error analyses above have some practical shortcomings. To apply the randomized absolute case one must have some a priori knowledge of the variance of the integrand. It must not larger than $\sigma_{\max}^2$ or the theory does not apply. If the variance of the integrand is much smaller than $\sigma_{\max}^2$, then the specified sample size, $\text{cost}^{\text{rnd abs}}(\varepsilon, \mathcal{F}^{\text{abs}}, \text{sMC})$ is much too large. The randomized relative case is even more problematic since determining whether $f \in \mathcal{F}^{\text{rel}}$ requires a priori knowledge of the ratio of the variance of the integrand to its squared mean. The randomized normalized case avoids the problem of requiring a priori knowledge of the integrand variance by including the variance of the integrand in the error criterion. However, still this does not give the practioner a clear idea of how large the error will be, only that it can be made small relative to the squre root of the variance of the integrand.

3. Probabilistic Error Analysis

Another approach that appears somewhat more stringent than the randomized error setting is the probabilistic error setting. In this case one wishes to be $(1-\alpha)100\%$ certain of achieving an error tolerance. Again there may be different cases:

(10a) probabilistic absolute

$$\text{err}^{\text{prob abs}}(n,\alpha,\mathcal{F},\text{sMC}) := \min\left\{\varepsilon : \sup_{f\in\mathcal{F}} \text{Prob}\left[|\hat{\mu}_n - \mu| \le \varepsilon\right] \ge 1-\alpha\right\},$$

(10b) probabilistic relative

$$\text{err}^{\text{prob rel}}(n,\alpha,\mathcal{F},\text{sMC}) := \min\left\{\varepsilon : \sup_{f\in\mathcal{F}} \text{Prob}\left[\left|\frac{\hat{\mu}_n - \mu}{\mu}\right| \le \varepsilon\right] \ge 1-\alpha\right\},$$

(10c) probabilistic normalized

$$\text{err}^{\text{prob nor}}(n,\alpha,\mathcal{F},\text{sMC}) := \min\left\{\varepsilon : \sup_{f\in\mathcal{F}} \text{Prob}\left[\left|\frac{\hat{\mu}_n - \mu}{\sigma}\right| \le \varepsilon\right] \ge 1-\alpha\right\},$$

Again one may define the minimum number of function values required to attain a given error tolerance as

$$\text{cost}^{\text{x}}(\varepsilon,\alpha,\mathcal{F},\text{sMC}) = \min\{n \in \mathbb{N}_0 : \text{err}^{\text{x}}(n,\alpha,\mathcal{F},\text{sMC}) \le \varepsilon\},$$

where $\text{x} \in \{\text{prob abs}, \text{prob rel}, \text{prob nor}\}$.

These error measures also have the practical shortcomings as described for the randomized error analysis. However, we will find ways to overcome them.

In practice, one often invokes the Central Limit Theorem to determine sample size. Given a significance level or uncertainty tolerance, α, one has

$$\text{Prob}\left[|\hat{\mu}_n - \mu| \le \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right] \approx 1-\alpha.$$

This directly leads to the *approximate* probabilistic result:

$$\text{(11a)} \qquad \text{cost}^{\text{prob abs}}(\varepsilon,\mathcal{F}^{\text{abs}},\text{sMC}) \approx N_G(\varepsilon/\sigma_{\max},\alpha), \qquad \text{where } N_G(\varepsilon,\alpha) := \left\lceil \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 \right\rceil,$$

$$\text{(11b)} \qquad \text{cost}^{\text{prob rel}}(\varepsilon,\mathcal{F}^{\text{rel}},\text{sMC}) \approx N_G(\varepsilon/a,\alpha),$$

$$\text{(11c)} \qquad \text{cost}^{\text{prob nor}}(\varepsilon,\mathcal{F}^{\text{nor}},\text{sMC}) \approx N_G(\varepsilon,\alpha).$$

The above is exact if the $Y_i = f(\boldsymbol{X}_i)$ are i.i.d. Gaussian, however, in general this result is only approximate.

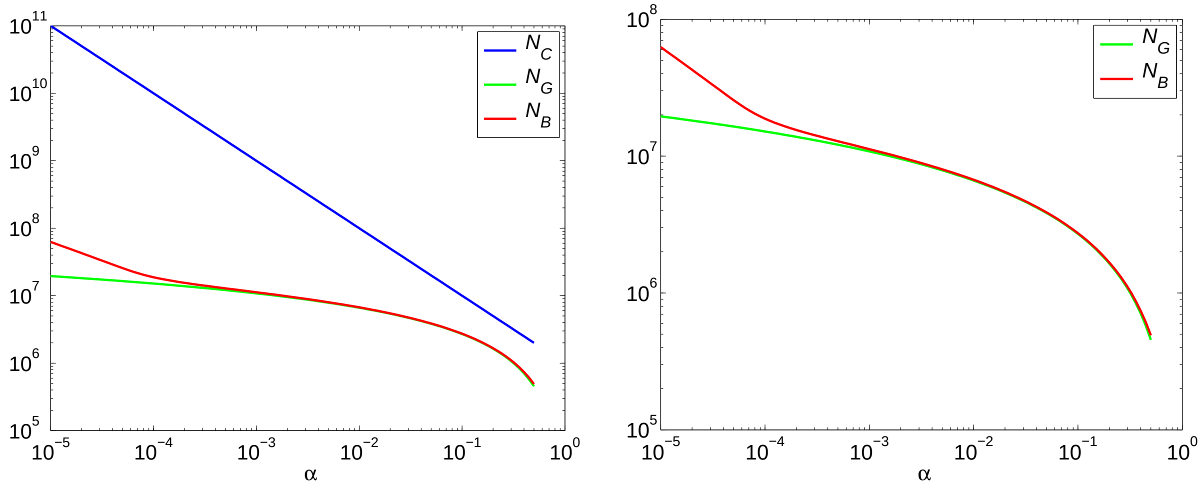

FIGURE 1. Comparison of $N_G(\varepsilon, \alpha)$, $N_C(\varepsilon, \alpha)$, and $N_B(\varepsilon, \alpha, \varrho)$ for $\varepsilon = 0.001$, and $\varrho = 5$.

Using Chebyshev's inequality (Theorem 3) yields an exact upper bound rather than approximate one. Choosing $Z = \hat{\mu}_n$ yields

$$\text{Prob}\left[|\hat{\mu}_n - \mu| < \frac{\sigma}{\sqrt{n\alpha}}\right] \geq 1 - \alpha.$$

This inequality leads to an *upper bound* on the sample size required to guarantee that the estimated mean is within the tolerance of the true answer with probability $1 - \alpha$:

$$\text{cost}^{\text{prob abs}}(\varepsilon, \mathcal{F}^{\text{abs}}, \text{sMC}) \leq N_C(\varepsilon/\sigma_{\max}, \alpha), \qquad \text{where } N_C(\varepsilon, \alpha) := \left\lceil \frac{1}{\alpha\varepsilon^2} \right\rceil, \tag{12a}$$

$$\text{cost}^{\text{prob rel}}(\varepsilon, \mathcal{F}^{\text{rel}}, \text{sMC}) \leq N_C(\varepsilon/a, \alpha), \tag{12b}$$

$$\text{cost}^{\text{prob nor}}(\varepsilon, \mathcal{F}^{\text{nor}}, \text{sMC}) \leq N_C(\varepsilon, \alpha). \tag{12c}$$

The sample size N_C is typically much larger than N_G, since $1/\alpha$ is much larger than $z_{\alpha/2}^2$ as $\alpha \to 0$. Figure 1 compares these two quantities.

A smaller sample size with a rigorous probabilistic bound can be found by invoking the non-uniform Berry-Esseen inequality (Theorem 4). This inequality makes strong assumptions on the distribution of $f(\boldsymbol{X})$, namely, a finite third moment:

$$\varrho := \frac{M_3}{\sigma^3} < \infty.$$

However, this assumption is weaker than the one that we intend to make below to justify the approximation of the variance by its sample quantity. Recalling that $Y_i = g(\boldsymbol{X}_i)$, $\mu = E(Y_i)$, and $\hat{\mu}_n = (Y_1 + \cdots + Y_n)/n$, it then follows by the non-uniform Berry-Esseen inequality,

$$\begin{aligned}\text{Prob}\left[|\hat{\mu}_n - \mu| < \frac{\sigma}{\sqrt{n}}x\right] &= \text{Prob}\left[\hat{\mu}_n - \mu < \frac{\sigma}{\sqrt{n}}x\right] - \text{Prob}\left[\hat{\mu}_n - \mu < -\frac{\sigma}{\sqrt{n}}x\right] \\ &\geq \left[\Phi(x) - A\frac{\varrho}{\sqrt{n}}(1+|x|)^{-3}\right] - \left[\Phi(-x) + A\frac{\varrho}{\sqrt{n}}(1+|x|)^{-3}\right] \\ &= 1 - 2\left(A\frac{\varrho}{\sqrt{n}}(1+|x|)^{-3} + \Phi(-x)\right).\end{aligned}$$

Letting $b = x/\sqrt{n}$, the probability of making a small error becomes

$$\text{Prob}[|\hat{\mu}_n - \mu| < b\sigma] \geq 1 - 2\left(A\frac{\varrho}{\sqrt{n}}\left(1 + b\sqrt{n}\right)^{-3} + \Phi\left(-b\sqrt{n}\right)\right) \geq 1 - \alpha,$$

provided that n is chosen to be larger than

$$N_B(b, \alpha, \varrho) := \min\left\{n \in \mathbb{N} : \Phi\left(-b\sqrt{n}\right) + \frac{A\varrho}{\sqrt{n}}\left(1 + b\sqrt{n}\right)^{-3} \leq \frac{\alpha}{2}\right\}. \tag{13}$$

This inequality leads to an *upper bound* on the work required to meet the probabilistic error criterion, but this time for a smaller set of functions. guarantee that the estimated mean is within the tolerance of the true answer with probability $1 - \alpha$:

$$\text{cost}^{\text{prob abs}}(\varepsilon, \mathcal{F}^{\text{abs}} \cap \mathcal{F}^{\text{third}}, \text{sMC}) \leq N_B(\varepsilon/\sigma_{\max}, \alpha, \varrho_{\max}), \tag{14a}$$

$$\text{cost}^{\text{prob rel}}(\varepsilon, \mathcal{F}^{\text{rel}} \cap \mathcal{F}^{\text{third}}, \text{sMC}) \leq N_B(\varepsilon/a, \alpha, \varrho_{\max}), \tag{14b}$$

$$\text{cost}^{\text{prob nor}}(\varepsilon, \mathcal{F}^{\text{nor}} \cap \mathcal{F}^{\text{third}}, \text{sMC}) \leq N_B(\varepsilon, \alpha, \varrho_{\max}), \tag{14c}$$

where

$$\mathcal{F}^{\text{third}} := \left\{f \in \mathcal{L}_3 : \frac{M_3(f)}{[\text{var}(f)]^{3/2}} \leq \rho_{\max}\right\}. \tag{14d}$$

Definition (13) may be re-written implicitly

$$N_B = \left\lceil \left(\frac{z_{\alpha/2 - A\varrho/(\sqrt{N_B}(1+b\sqrt{N_B})^3)}}{b}\right)^2 \right\rceil. \tag{15}$$

The definition of N_B implies that

$$N_G(b, \alpha) \leq N_B(b, \alpha, \varrho) \leq \min_{0 \leq \theta \leq 1}\left\{\max\left[\sqrt{\frac{2A\rho}{\theta\alpha b^3}}, N_G(b, (1-\theta)\alpha)\right]\right\}.$$

As shown in Figure 1, N_B is close to N_G for moderate α, but N_B may be significantly larger for very small α.

4. Adaptive Monte Carlo

The Chebyshev and the Berry-Esseen inequalities provide upper bounds on the sample size required by the simple Monte Carlo algorithm to attain the desired error tolerance under the probabilistic error criterion, (12) and (14), respectively. However, as mentioned earlier, in the case of the absolute and relative error criterion, the classes of integrands are defined in terms of quantities, such as $\sigma^2 = \text{var}(f)$, which are typically unknown. For the case of normalized error, the error criterion is defined in terms of $\text{var}(f)$.

In practice one typically uses observed function values observed to approximate σ^2 by the sample variance, as follows:

$$\hat{v}_n = \frac{1}{n-1}\sum_{i=1}^{n}[f(\boldsymbol{X}_i) - \hat{\mu}_n]^2. \tag{16}$$

This means that we now have an *adaptive* algorithm. One might choose an initial sample of size n_0, and use it to estimate σ^2 by $\hat{v}_{n_0}$. Then one chooses an *independent* sample of size $n = N_C(\varepsilon/\sqrt{\hat{v}_{n_0}}, \alpha)$, or $n = N_B(\varepsilon/\sqrt{\hat{v}_{n_0}}, \alpha, \varrho_{\max})$ to compute $\hat{\mu}_n$ the final estimate of μ.

Unfortunately, once we approximate σ^2 by $\hat{v}_n$, we again have inexact results. However, they can be made exact by using Cantelli's inequality (Theorem 6) and the variance of $\hat{v}_n$ in Theorem 5. Proposition 7 implies that

$$\text{Prob}\left[\frac{\hat{v}_n}{1-\sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}} > \sigma^2\right] \geq 1-\alpha,$$

where

$$\kappa := \text{kurt}(f) = \frac{M_4(f)}{\text{var}^2(f)} = \frac{M_4(f)}{\sigma^4(f)} \geq 1 \qquad \forall f \in \boldsymbol{\mathcal{L}}_4 \tag{17}$$

denotes the *kurtosis*. Note that the kurtosis of a function is independent of scale. Thus, it follows that

$$\text{Prob}\left[\hat{\sigma}_{\text{up}}(\hat{v}_n, n, \alpha, \kappa) > \sigma\right] \geq 1-\alpha, \tag{18a}$$

$$\text{where } \hat{\sigma}^2_{\text{up}}(\hat{v}_n, n, \alpha, \kappa) = \frac{\hat{v}_n}{1-\sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}}, \tag{18b}$$

provided that

$$\begin{aligned} 1 &> \left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right) \\ \frac{\alpha n}{1-\alpha} &> \kappa - \frac{n-3}{n-1} \\ \kappa &< \frac{\alpha n}{1-\alpha} + \frac{n-3}{n-1} =: \kappa_{\text{poss}}(\alpha, n). \end{aligned}$$

From another perspective, if we want the amplification factor to be smaller than L, i.e.,

$$\frac{1}{1-\sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}} \leq L,$$

then n must be chosen to satisfy

$$\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right) \le \frac{L-1}{L},$$

$$M\left[(\kappa-1) + \frac{2}{n-1}\right] \le n, \qquad \text{where } M = \left(\frac{1-\alpha}{\alpha}\right)\left(\frac{L}{L-1}\right),$$

$$(n-1)^2 - [(\kappa-1)M-1](n-1) - 2M \ge 0,$$

$$n \ge 1 + \frac{1}{2}\left\{(\kappa-1)M - 1 + \sqrt{[(\kappa-1)M-1]^2 + 8M}\right\}.$$

Thus, for example, even in the best case scenario of $\kappa = 1$, the smallest possible value, we have $n \ge [1+\sqrt{1+8M}]/2$. For $L = 2$ and $\alpha = 0.05$ this translates into $M = 38$ and $n \ge 9.2$.

On the other hand, suppose that we choose

$$\hat{\sigma}^2_{\text{up}} = L\hat{v}_n$$

where L is a *fudge factor*. This means that we are implicitly assuming the kurtosis to satisfy

$$\frac{1}{1 - \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}} \le L,$$

$$\frac{1}{L} \le 1 - \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)},$$

$$\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right) \le \left(1 - \frac{1}{L}\right)^2 = \left(\frac{L-1}{L}\right)^2,$$

$$\kappa \le \frac{n-3}{n-1} + \left(\frac{1-\alpha}{\alpha n}\right)\left(\frac{L-1}{L}\right)^2.$$

From still another perspective, for a given n and κ, one must require α large enough, namely

$$\frac{1}{n}\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha}\right) < 1,$$

$$\alpha > \frac{1}{1 + \frac{n}{\kappa - \frac{n-3}{n-1}}}.$$

Figure 2 shows how much the sample variance needs to be inflated to be confident that the true variance is not underestimated. Notice that for $n = 30$, the common rule of thumb for applying the central limit theorem, even $\alpha = 0.1$ is not possible.

Theorem 1. *For a given positive constant, $\kappa_{\max}$, define the set of functions with bounded fourth moments:*

$$\mathcal{F}^{\text{kurt}} = \{f \in \mathcal{L}_4 : \text{kurt}(f) = \kappa \le \kappa_{\max}\},$$

where the kurtosis of the function is defined in (17). *Suppose that one has an error tolerance, ε, and an uncertainty tolerance, α. Let $\alpha_1 = 1 - \sqrt{1-\alpha}$. Pick any $n_0 > 1$ satisfying*

$$\kappa_{\max} < \kappa_{\text{poss}}(\alpha_1, n_0) = \frac{n_0\alpha_1}{1-\alpha_1} + \frac{n_0 - 3}{n_0 - 1},$$

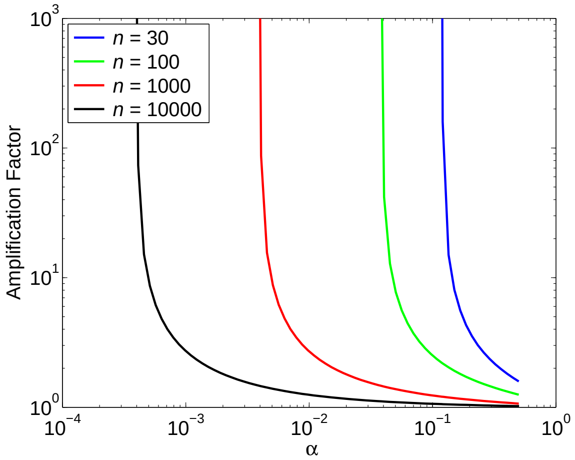

FIGURE 2. The amplification factor, $\hat{\sigma}^2_{\text{up}}/\sigma^2$, as a function of α for various n and for $\kappa = 5$.

and compute the sample variance, $\hat{v}_{n_0}$ of a simple random sample of size n_0. Use this to compute $\hat{\sigma}^2_{\text{up}} = \hat{\sigma}^2_{\text{up}}(\hat{v}_{n_0}, n_0, \alpha_1, \kappa_{\max})$ by (18). Next choose an independent random sample of size

$$n = \min\left(N_C(\varepsilon/\hat{\sigma}_{\text{up}}, \alpha_1), N_B(\varepsilon/\hat{\sigma}_{\text{up}}, \alpha_1, \kappa_{\max}^{3/4})\right)$$

and compute $\hat{\mu}_n$, the simple Monte Carlo estimator of μ. Here N_C is defined in (12) and N_B is defined in (13). A probabilistic error bound is given by

$$\text{Prob}\left[|\hat{\mu}_n - \mu| \le \epsilon\right] \ge 1 - \alpha.$$

Proof. By (18) it follows that $\hat{\sigma}_{\text{up}}(\hat{v}_{n_0}, n_0, \alpha_1, \kappa) \ge \sigma$ with probability $1-\alpha_1$. By (12) and (14) it follows that $\text{Prob}\left[|\hat{\mu}_n - \mu| \le \epsilon\right] \ge 1 - \alpha_1$, provided that $\hat{\sigma}_{\text{up}} \ge \sigma$. Thus, the probability that both of these events happen, is at least $(1-\alpha_1)^2 = 1 - \alpha$. □

The sample size of this algorithm is now a random variable, and so the cost is defined probabilistically. Define the cost of an algorithm as the $1-\beta$ quantile of the total number of function evaluations. Furthermore, the cost now depends not only on the space of functions, but also on the variance of the integrand, which is stated explicitly:

$$\text{cost}(\varepsilon, \sigma^2, \mathcal{F}^{\text{kurt}}, \text{aMC}) := \sup_{\substack{f \in \mathcal{F}^{\text{kurt}} \\ \text{var}(f) \le \sigma^2}} \min\left\{M : \text{Prob}(n_0 + n \le M) \ge 1 - \beta\right\}. \tag{19}$$

From form Proposition 7 it follows that

$$\begin{aligned} 1 - \beta &\le \text{Prob}\left[\hat{v}_n < \sigma^2\left\{1 + \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\beta}{\beta n}\right)}\right\}\right] \\ &= \text{Prob}\left[\frac{\hat{v}_n}{1 - \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha_1}{\alpha_1 n}\right)}} < \sigma^2\left\{\frac{1 + \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\beta}{\beta n}\right)}}{1 - \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha_1}{\alpha_1 n}\right)}}\right\}\right] \\ &= \text{Prob}\left[\hat{\sigma}^2_{\text{up}}(\hat{v}_n, n, \alpha_1, \kappa_{\max}) < \sigma^2 \gamma^2(n, \alpha_1, \beta, \kappa_{\max})\right], \end{aligned}$$

where,

$$\gamma^2(n, \alpha_1, \beta, \kappa_{\max}) := \frac{1 + \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\beta}{\beta n}\right)}}{1 - \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha_1}{\alpha_1 n}\right)}} > 1.$$

Since $N_C(\cdot, \alpha_1)$ and $N_B(\cdot, \alpha_1, \kappa_{\max}^{3/4})$ are decreasing functions, it follows that

$$\begin{aligned}
(20) \quad \operatorname{cost}(\varepsilon, \sigma^2, \mathcal{F}^{\text{kurt}}, \text{aMC}) &= \sup_{\substack{f \in \mathcal{F}^{\text{kurt}} \\ \operatorname{var}(f) \le \sigma^2}} \min\left\{M : \operatorname{Prob}(n_0 + n \le M) \ge 1 - \beta\right\} \\
&= n_0 + \min\left\{M : \operatorname{Prob}\left(\min\left(N_C(\varepsilon/\hat{\sigma}_{\text{up}}, \alpha_1), N_B(\varepsilon/\hat{\sigma}_{\text{up}}, \alpha_1, \kappa_{\max}^{3/4})\right) \le M\right) \ge 1 - \beta\right\}]] \\
&\le n_0 + \min\left(N_C(\varepsilon/(\sigma\gamma(n_0, \alpha_1, \beta, \kappa_{\max})), \alpha_1), N_B(\varepsilon/(\sigma\gamma(n_0, \alpha_1, \beta, \kappa_{\max})), \alpha_1, \kappa_{\max}^{3/4})\right).
\end{aligned}$$

Theorem 2. *The algorithm described in Theorem 1 has a probabilistic cost given by* (20).

The key factors that determine $\operatorname{cost}(\varepsilon, \sigma^2, \mathcal{F}^{\text{kurt}}, \text{aMC})$ are ε, the error tolerance, and σ^2, the variance of the integrand. The cost is roughly proportional to $\sigma^2\epsilon^{-2}$. For the set of integrands $\mathcal{F}^{\text{kurt}}$ the variance, $\operatorname{var}(f)$ is unbounded. Thus, the cost is not bounded, however, it does seem to behave as expected as a function of the variance of the integrand. As mentioned before, this is actually an advantage of this analysis. One need not make any assumptions about the variance of the integrand, only about the kurtosis, which is unchanged when the integrand is multiplied by an arbitrary constant.

5. Example

Consider the case of the uniform probability distribution on $[0, 1]$, i.e., $\rho = 1$. Define

$$(21) \qquad f(x) = \begin{cases} 1 + \sigma\sqrt{\frac{1-p}{p}}, & 0 \le x \le p, \\ 1 - \sigma\sqrt{\frac{p}{1-p}}, & p < x \le 1, \end{cases}$$

where p and σ are parameters, with $0 < p < 1$. Note that

$$\begin{aligned}
\mu &= \int_0^1 g(x)\,\mathrm{d}x = 1 \\
\operatorname{var}(g) &= \int_0^1 [g(x) - \mu]^2\,\mathrm{d}x = \sigma^2\frac{1-p}{p}p + \sigma^2\frac{p}{1-p}(1-p) = \sigma^2, \\
\kappa = \operatorname{kurt}(g) &= \frac{1}{\sigma^4}\int_0^1 [g(x) - \mu]^4\,\mathrm{d}x = \left(\frac{1-p}{p}\right)^2 p + \left(\frac{p}{1-p}\right)^2 (1-p) \\
&= \frac{(1-p)^3 + p^3}{p(1-p)} = \frac{1 - 3p + 3p^2}{p(1-p)} = \frac{1}{p(1-p)} - 3.
\end{aligned}$$

Note that κ ranges from a minimum of 1, when $p = 1/2$ to a maximum of ∞ when $p = 0, 1$.

Figure 3 shows the empirical distribution of the normalized error $|\mu - \hat{\mu}_n|/\epsilon$, using 1000 replications for a range of values of p. As can be seen in this figure and in Table 1, the adaptive Monte Carlo method does poorly for very small values of p, which correspond to vary large values of the kurtosis. However, even for values of the kurtosis above $\kappa_{\max} = 3.2$ used in this example, the chance of meeting the error tolerance may be quite high.

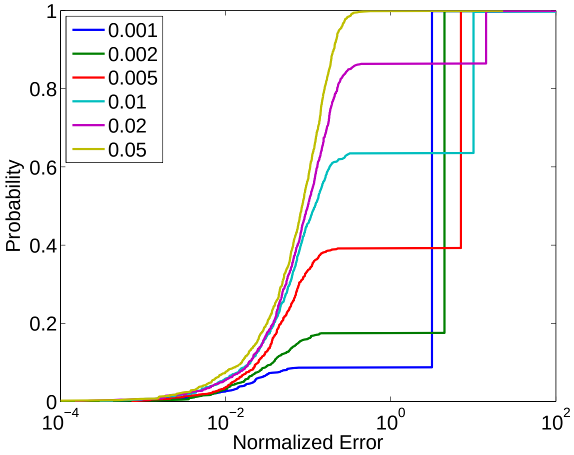

FIGURE 3. Empirical distribution function of $|\mu - \hat{\mu}_n| / \epsilon$ for example (21) with $\mu = \sigma = 1$, $n_0 = 100$, $\kappa_{\max} = 3.2$, $\varepsilon = 0.01$, and $p = 0.001, 0.002, 0.005, 0.01, 0.02, 0.05$ using the algorithm in Theorem 1.

TABLE 1. Kurtosis and probability of meeting the error tolerance for different values of p.

p	0.001	0.002	0.005	0.01	0.02	0.05
κ	998.00	498.00	198.01	98.01	48.02	18.05
$\text{Prob}(\lvert\mu - \hat{\mu}_n\rvert \le \varepsilon)$	8.70%	17.50%	39.20%	63.50%	86.40%	99.90%

6. QUESTIONS

Here are some questions that suggest themselves:

- Is this analysis above known already? Is this the typical probabilistic setting? Is it better to look at a randomized setting where one considers the expected value of the error?
- Can this type of analysis be extended to randomized *quasi-Monte Carlo* for finite dimension, d? Infinite dimension? In this latter case one needs some multilevel algorithm, but the specification of the levels perhaps could be deduced from the data. One might also consider a case where the coordinate weights were not known a priori but needed to be estimated.
- Is there already this kind of information-based complexity analysis where the number of operations is bounded above or below by the unknown scale of the problem (in this case the variance). The IBC I know assumes that the scale is fixed, e.g., the function has variance one, norm one, etc. Here we allow arbitrary scale, but do make assumptions on the nastiness (kurtosis).
- Are there better inequalities than Chebyshev's inequality or the Berry-Esseen inequality that apply when Z is the sum of i.i.d. random variables? Some of the better known ones, like Hoeffding's inequality assume boundedness, which we cannot presume here.

Appendix of Useful Theorems

Theorem 3 (Chebyshev's Inequality)**.** *Let Z be any random variable with mean μ and variance σ^2. Then for all $\alpha > 0$, Chebyshev's inequality states that*

$$\operatorname{Prob}\left[|Z-\mu| \ge \frac{\sigma}{\sqrt{\alpha}}\right] \le \alpha, \qquad \operatorname{Prob}\left[|Z-\mu| < \frac{\sigma}{\sqrt{\alpha}}\right] \ge 1-\alpha.$$

Proof. To prove Chebyshev's inequality note that

$$\sigma^2 = E[|Z-\mu|^2] \ge \frac{\sigma^2}{\alpha} \operatorname{Prob}\left[|Z-\mu| \ge \frac{\sigma}{\sqrt{\alpha}}\right],$$

and then divide both sides by σ^2/α. □

The following theorem comes from (Petrov, 1995, Theorem 5.16, p. 168)

Theorem 4 (Non-uniform Berry-Esseen Inequality)**.** *Let $Y_1, \ldots, Y_n$ be i.i.d. random variables. Suppose that*

$$\mu = E(Y_i), \quad \operatorname{var}(Y_i) = \sigma^2 > 0, \quad \varrho = \frac{E\,|Y_i-\mu|^3}{\sigma^3} < \infty.$$

Then

$$\left|\operatorname{Prob}\left[\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(Y_i-\mu) < x\right] - \Phi(x)\right| \le \frac{A\varrho}{\sqrt{n}}(1+|x|)^{-3}.$$

for all x, where Φ is the cumulative distribution function of the standard normal random variable, and A is some number satisfying $0.4097 \le A \le 0.5600$.

Theorem 5. *Let $\hat{v}_n$ be the sample variance as defined in* (16)*. It's variance is*

$$\operatorname{var}(\hat{v}_n^2) = \frac{\sigma^4}{n}\left(\kappa - \frac{n-3}{n-1}\right),$$

where $\kappa :=$ kurt$(g) = M_4(g)/\sigma^4(g)$ *denotes the* kurtosis.

Proof. The sample variance has mean σ^2/n. To facilitate the derivation, let $Y_i = g(X_i) - \mu$.

$$\begin{aligned}
\hat{v}_n &= \frac{1}{n-1}\sum_{i=1}^{n}\left[Y_i - \left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right)\right]^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} Y_i^2 - \sum_{j,k=1}^{n} Y_j Y_k\right] \\
\hat{v}_n^2 &= \frac{1}{n^2(n-1)^2}\left[n^2\sum_{i,j=1}^{n} Y_i^2 Y_j^2 - 2n\sum_{i,j,k=1}^{n} Y_i^2 Y_j Y_k + \sum_{i,j,k,l=1}^{n} Y_i Y_j Y_k Y_l\right] \\
E[Y_i^2 Y_j^2] &= \begin{cases} M_4, & i = j, \\ \sigma^4, & i \neq j, \end{cases} \\
\sum_{i,j=1}^{n} E[Y_i^2 Y_j^2] &= nM_4 + n(n-1)\sigma^4, \\
E[Y_i^2 Y_j Y_k] &= \begin{cases} M_4, & i = j = k, \\ \sigma^4, & i \neq j, j = k, \\ 0, & j \neq k, \end{cases} \\
\sum_{i,j,k=1}^{n} E[Y_i^2 Y_j Y_k] &= nM_4 + n(n-1)\sigma^4 \\
E[Y_i Y_j Y_k Y_l] &= \begin{cases} M_4, & i = j = k = l, \\ \sigma^4, & i, j, k, l \text{ have 2 distinct values}, \\ 0, & \text{otherwise}, \end{cases} \\
\sum_{i,j,k,l=1}^{n} E[Y_i Y_j Y_k Y_l] &= nM_4 + 3n(n-1)\sigma^4 \\
E[\hat{v}_n^2] &= \frac{n^3[M_4 + (n-1)\sigma^4] - 2n^2[M_4 + (n-1)\sigma^4] + n[M_4 + 3(n-1)\sigma^4]}{n^2(n-1)^2} \\
&= \frac{(n-1)M_4 + (n^2 - 2n + 3)\sigma^4}{n(n-1)} \\
\mathrm{var}(\hat{v}_n^2) &= E[\hat{v}_n^2] - [E(\hat{v}_n)]^2 = \frac{(n-1)M_4 + (n^2 - 2n + 3)\sigma^4}{n(n-1)} - \sigma^4 \\
&= \frac{(n-1)M_4 + (-n+3)\sigma^4}{n(n-1)} = \frac{1}{n}\left(M_4 - \frac{n-3}{n-1}\sigma^4\right) = \frac{\sigma^4}{n}\left(\kappa - \frac{n-3}{n-1}\right).
\end{aligned}$$

□

Theorem 6 (Single tailed Cantelli's inequality)**.** *Let Z be any random variable with mean μ and finite variance σ^2. For any $a \geq 0$, it follows that:*

$$\mathrm{Prob}[Z - \mu \geq a] \leq \frac{\sigma^2}{a^2 + \sigma^2}.$$

Proof. Define the random variable

$$S = \mathrm{sign}(Z - \mu - a) = \begin{cases} 1, & Z - \mu \geq a, \\ -1, & Z - \mu < a. \end{cases}$$

From conditional probability it is known that

$$\begin{aligned}\sigma^2 &= \operatorname{var}(Z-\mu) = E[\operatorname{var}(Z-\mu|S)] + \operatorname{var}[E(Z-\mu|S)]\\ &\geq \operatorname{var}[E(Z-\mu|S)] = E[\{E(Z-\mu|S)\}^2] - [E\{E(Z-\mu|S)\}]^2 = E[\{E(Z-\mu|S)\}^2]\end{aligned}$$

Since $E(Z-\mu)=0$, it follows that

$$0 = E[E(Z-\mu|S)] = E(X|S=1)\operatorname{Prob}(Z-\mu \geq a) + E(X|S=-1)\operatorname{Prob}(Z-\mu < a).$$

Also, it is clear that $E(Z-\mu|S=1)\geq a$, which implies that

$$[E(Z-\mu|S=-1)]^2 = \left[\frac{E(Z-\mu|S=1)\operatorname{Prob}(Z-\mu\geq a)}{\operatorname{Prob}(Z-\mu<a)}\right]^2 \geq \left[\frac{a\operatorname{Prob}(Z-\mu\geq a)}{\operatorname{Prob}(Z-\mu<a)}\right]^2$$

Combining these results together yields

$$\begin{aligned}\sigma^2 &\geq E[\{E(Z-\mu|S)\}^2]\\ &= \{E(Z-\mu|S=1)\}^2\operatorname{Prob}(Z-\mu\geq a) + \{E(Z-\mu|S=-1)\}^2\operatorname{Prob}(Z-\mu<a)\\ &\geq a^2\operatorname{Prob}(Z-\mu\geq a) + \left[\frac{a\operatorname{Prob}(Z-\mu\geq a)}{\operatorname{Prob}(Z-\mu<a)}\right]^2\operatorname{Prob}(Z-\mu<a)\\ &= a^2\left[\frac{\operatorname{Prob}(Z-\mu\geq a)}{\operatorname{Prob}(Z-\mu<a)}\right] = a^2\left[\frac{\operatorname{Prob}(Z-\mu\geq a)}{1-\operatorname{Prob}(Z-\mu\geq a)}\right]\end{aligned}$$

Solving this inequality for $\operatorname{Prob}(Z-\mu\geq a)$ completes the proof. □

Proposition 7. *Let $\hat{v}_n$ be the sample variance of a function g as defined in* (16)*, and let $\kappa = \operatorname{kurt}(g)$. Then*

$$\operatorname{Prob}\left[\hat{v}_n < \sigma^2\left\{1+\sqrt{\left(\kappa-\frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right\}\right] \geq 1-\alpha, \tag{22a}$$

$$\operatorname{Prob}\left[\hat{v}_n > \sigma^2\left\{1-\sqrt{\left(\kappa-\frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right\}\right] \geq 1-\alpha. \tag{22b}$$

Proof. Choosing

$$a = \sigma^2\sqrt{\left(\kappa-\frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)} > 0,$$

we know from Cantelli's inequality (Theorem 6) that

$$\operatorname{Prob}[\hat{v}_n - \sigma^2 \geq a] \leq \frac{\operatorname{var}(\hat{v}_n)}{a^2+\operatorname{var}(\hat{v}_n)}$$

$$\begin{aligned}
\operatorname{Prob}\left[\hat{v}_n - \sigma^2 \ge \sigma^2 \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right] &= \operatorname{Prob}\left[\hat{v}_n - \sigma^2 \ge a\right] \\
&\le \frac{\operatorname{var}(\hat{v}_n)}{a^2 + \operatorname{var}(\hat{v}_n)} \\
&= \frac{\frac{\sigma^4}{n}\left(\kappa - \frac{n-3}{n-1}\right)}{\frac{\sigma^4}{n}\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha}\right) + \frac{\sigma^4}{n}\left(\kappa - \frac{n-3}{n-1}\right)} \\
&= \frac{1}{\left(\frac{1-\alpha}{\alpha}\right) + 1} = \alpha.
\end{aligned}$$

Then (22a) follows directly. By a similar argument.

$$\begin{aligned}
\operatorname{Prob}\left[\hat{v}_n - \sigma^2 \le -\sigma^2 \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right] &= \operatorname{Prob}\left[\hat{v}_n - \sigma^2 \le -a\right] \\
&= \operatorname{Prob}\left[(-\hat{v}_n) - (-\sigma^2) \ge a\right] \\
&\le \frac{\operatorname{var}(-\hat{v}_n)}{a^2 + \operatorname{var}(-\hat{v}_n)} = \frac{\operatorname{var}(\hat{v}_n)}{a^2 + \operatorname{var}(\hat{v}_n)} \\
&= \frac{\frac{\sigma^4}{n}\left(\kappa - \frac{n-3}{n-1}\right)}{\frac{\sigma^4}{n}\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha}\right) + \frac{\sigma^4}{n}\left(\kappa - \frac{n-3}{n-1}\right)} \\
&= \frac{1}{\left(\frac{1-\alpha}{\alpha}\right) + 1} = \alpha.
\end{aligned}$$

Thus, (22b) follows as well. □

References

Petrov VV (1995) Limit Theorems of Probability Theory:Sequences of Independent Random Variables. Clarendon Press, Oxford

Room E1-208, Department of Applied Mathematics, Illinois Institute of Technology, 10 W. 32nd St., Chicago, IL 60616

Room E1-208, Department of Applied Mathematics, Illinois Institute of Technology, 10 W. 32nd St., Chicago, IL 60616

School of Mathematics and Statistics, Lanzhou University, Lanzhou City, Gansu, China 730000