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. Non-Adaptive Monte Carlo

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

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

where $f : \mathbb{R}^d \to [0, \infty)$ is a probability density function. The $\mathcal{L}_{p,f}$ norm of g is defined by

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

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

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

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

$$M_p(g) := \int_{\mathbb{R}^d} |g(\boldsymbol{x}) - \mu|^p \, f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \qquad p \ge 1 \tag{2}$$

denotes the centered absolute moments of g, then

$$M_q \le M_p^{q/p} \quad \text{and} \quad M_p < \infty \implies M_q < \infty \qquad \text{for } 1 \le q \le p. \tag{3}$$

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be independent and identically distributed random variables with the probability density function f, briefly, $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ i.i.d. $\sim f$. Then the simple Monte Carlo estimator of the mean, is

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{X}_i). \tag{4}$$

An important question that springs to mind is how to choose the right sample size, n, to achieve a specified tolerance, ϵ. The random estimator, $\hat{\mu}_n$ has mean μ and variance

$$\operatorname{var}(\hat{\mu}_n) = \frac{\sigma^2}{n},$$

where σ^2 is the variance of the function g:

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

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

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

Thus one has the *approximate* probabilistic result

$$\operatorname{Prob}\left[|\hat{\mu}_n - \mu| \le \epsilon\right] \approx 1 - \alpha \qquad \text{for } n = N_N(\epsilon, \alpha, \sigma^2) := \left\lceil \left(\frac{z_{\alpha/2}\sigma}{\epsilon}\right)^2 \right\rceil. \tag{6}$$

The above is exact if $g(\boldsymbol{X}_i)$ are i.i.d. normal, however, in general this result is only approximate and relies on the unknown σ. The above observations are formalized in the proposition below.

Proposition 1. *For a given positive constant, $\sigma_{\max}$, define*

$$\mathcal{G}_{2,N,\sigma_{\max}} = \{g \in \mathcal{L}_{2,f} : g(\boldsymbol{X}) \sim N(\mu, \sigma^2) \text{ for } \boldsymbol{X} \sim f, \ \sigma^2 \le \sigma_{\max}^2\},$$

where the variance of the function is defined in (5)*. If for a given error tolerance, ϵ, and an uncertainty tolerance, α, then a probabilistic error bound is given by*

$$\operatorname{Prob}\left[|\hat{\mu}_n - \mu| \le \epsilon\right] \ge 1 - \alpha \qquad \text{for } n \ge N_N(\epsilon, \alpha, \sigma_{\max}^2).$$

Note that the set $\mathcal{G}_{2,N,\sigma_{\max}}$ assumes that $g(\boldsymbol{X})$ is normal, and that there is a priori knowledge about the variance of g. The first of these drawbacks may be removed by turning to Chebyshev's inequality.

Using Chebyshev's inequality (Theorem 6) yields an exact upper bound rather than approximate one. Choosing $Z = \hat{\mu}_n$ yields

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

Thus, a proper choice of sample size guarantees that the estimate is within the tolerance of the true answer with probability $1 - \alpha$:

$$\operatorname{Prob}\left[|\hat{\mu}_n - \mu| < \epsilon\right] \ge 1 - \alpha \qquad \text{for } n = N_C(\epsilon, \alpha, \sigma^2) := \left\lceil \frac{\sigma^2}{\alpha\epsilon^2} \right\rceil. \tag{7}$$

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

Proposition 2. *For a given positive constant, $\sigma_{\max}$, define the set of functions*

$$\mathcal{G}_{2,C,\sigma_{\max}} = \{g \in \mathcal{L}_{2,f} : \operatorname{var}(g) = \sigma^2 \le \sigma_{\max}^2\},$$

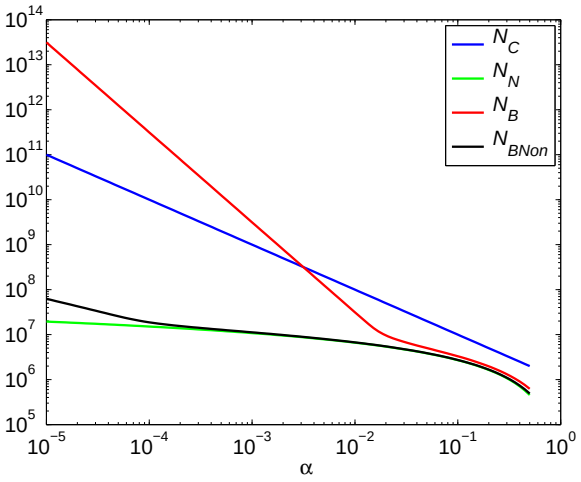

FIGURE 1. Comparison of N_N, N_C, and N_B for $\epsilon/\sigma = 0.001$, and $\rho = 5$.

where the variance of the function is defined in (5)*. If for a given error tolerance,* ϵ*, and an uncertainty tolerance,* α*, then a probabilistic error bound is given by*

$$\text{Prob}\left[|\hat{\mu}_n - \mu| \le \epsilon\right] \ge 1 - \alpha \qquad \textit{for } n \ge N_C(\epsilon, \alpha, \sigma^2_{\max}).$$

A smaller sample size with a rigorous probabilistic bound can be found by invoking the Berry-Esseen inequality (Theorem 7). This inequality makes strong assumptions on the distribution of $g(\boldsymbol{X})$, namely, a finite third moment:

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

Letting $Y_i = g(\boldsymbol{X}_i)$, it follows that $\mu = E(Y_i)$ and $\hat{\mu}_n = (Y_1 + \cdots + Y_n)/n$, and, furthermore, by the Berry-Esseen inequality,

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

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

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

provided that n is chosen to be larger than $N_B(\epsilon, \alpha, \sigma^2, \varrho)$, which is defined as the smallest integer satisfying

$$\Phi\left(-\frac{\sqrt{N_B}\epsilon}{\sigma}\right) + \frac{A\varrho}{\sqrt{N_B}} \le \frac{\alpha}{2}. \tag{8}$$

This equation may be re-written as the implicit equation

$$N_B = \left\lceil \left(\frac{z_{\alpha/2 - A\varrho/\sqrt{N_B}} \times \sigma}{\epsilon} \right)^2 \right\rceil. \tag{9}$$

The definition of N_B implies that

$$\max\left[\left(\frac{2A\rho}{\alpha}\right)^2, N_N(\epsilon,\alpha,\sigma^2)\right] \le N_B \le \max\left[\left(\frac{2A\rho}{\theta\alpha}\right)^2, N_N(\epsilon,(1-\theta)\alpha,\sigma^2)\right], \quad 0 \le \theta \le 1.$$

As shown in Figure 1, N_B is close to N_N for moderate α, but N_B may be even larger than N_C for very small α.

Proposition 3. *For given positive constants, $\sigma_{\max}$ and $\rho_{\max}$, define the set of functions*

$$\mathcal{G}_{3,B,\sigma_{\max},\varrho_{\max}} = \{g \in \mathcal{L}_{3,f} : \operatorname{var}(g) = \sigma^2 \le \sigma_{\max}^2,\ M_3(g)/\sigma^3 \le \varrho_{\max}\},$$

where the variance of the function is defined in (5)*, and M_3 is defined in* (2)*. If for a given error tolerance, ϵ, an uncertainty tolerance, α, then a probabilistic error bound is given by*

$$\operatorname{Prob}\left[|\hat{\mu}_n - \mu| \le \epsilon\right] \ge 1 - \alpha \qquad \text{for } n \ge N_B(\epsilon,\alpha,\sigma_{\max}^2,\varrho_{\max}),$$

where N_B is defined in (8)*.*

2. Adaptive Monte Carlo

Although the probabilitistic error results in Propositions 2 and 3 based on the Chebyshev and Berry-Esseen inequalities are exact for general distributions, the classes of integrands are defined in terms of the typically unknown $\sigma^2 = \operatorname{var}(g)$. This is why in practice one typically uses observed function values observed to approximate the σ^2 by the sample variance, as follows:

$$\hat{v}_n = \frac{1}{n-1}\sum_{i=1}^{n}[g(\boldsymbol{X}_i) - \hat{\mu}_n]^2. \tag{10}$$

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 $n = N_N(\epsilon,\alpha,\hat{v}_{n_0})$, $n = N_C(\epsilon,\alpha,\hat{v}_{n_0})$, or $n = N_B(\epsilon,\alpha,\hat{v}_{n_0},\rho_{\max})$ or 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 9) and the variance of $\hat{v}_n$ in Theorem 8. Proposition 10 implies that

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

where

$$\kappa := \operatorname{kurt}(g) = \frac{M_4(g)}{\operatorname{var}^2(g)} = \frac{M_4(g)}{\sigma^4(g)} \ge 1 \tag{11}$$

denotes the *kurtosis*. Thus,

$$\text{Prob}\left[\hat{\sigma}_{\text{up}}(\hat{v}_n, n, \alpha, \kappa) > \sigma\right] \geq 1 - \alpha, \tag{12a}$$

$$\text{where } \hat{\sigma}^2_{\text{up}}(\hat{v}_n, n, \alpha, \kappa) = \frac{\hat{v}_n}{1 - \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}}, \tag{12b}$$

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

$$\begin{gathered} \left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right) \leq \frac{L-1}{L}, \\ M\left[(\kappa - 1) + \frac{2}{n-1}\right] \leq n, \qquad \text{where } M = \left(\frac{1-\alpha}{\alpha}\right)\left(\frac{L}{L-1}\right), \\ (n-1)^2 - [(\kappa-1)M - 1](n-1) - 2M \geq 0, \\ n \geq 1 + \frac{1}{2}\left\{(\kappa-1)M - 1 + \sqrt{[(\kappa-1)M-1]^2 + 8M}\right\}. \end{gathered}$$

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

Theorem 4. *For a given positive constant, $\kappa_{\max}$, define the set of functions with bounded fourth moments:*

$$\mathcal{G}_{4,\kappa_{\max}} = \{g \in \mathcal{L}_{4,f} : \text{kurt}(g) = \kappa \leq \kappa_{\max}\},$$

where the kurtosis of the function is defined in (11)*. 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*

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

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 (12)*. Next choose an independent random sample of size*

$$n = \min\left(N_C(\epsilon, \alpha_1, \hat{\sigma}^2_{\text{up}}), N_B(\epsilon, \alpha_1, \hat{\sigma}^2_{\text{up}}, \kappa^{3/4}_{\max})\right)$$

and compute $\hat{\mu}_n$, the simple Monte Carlo estimator of μ. Here N_C is defined in (7) *and N_B is defined in* (8)*. A probabilistic error bound is given by*

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

Proof. By (12) it follows that $\hat{\sigma}_{\text{up}}(\hat{v}_{n_0}, n_0, \alpha_1, \kappa) \geq \sigma$ with probability $1-\alpha_1$. By Propositions 2 and 3 it follows that $\text{Prob}\left[|\hat{\mu}_n - \mu| \leq \epsilon\right] \geq 1 - \alpha_1$, provided that $\hat{\sigma}_{\text{up}} \geq \sigma$. Thus, the probability that both of these events happen, is at least $(1-\alpha_1)^2 = 1 - \alpha$. □

The cost of this algorithm can be evaluated in a probabilistic way also. Define the cost of an algorithm as the $1-\beta$ quantile of the total number of function evaluations:

$$\text{cost}(\hat{\mu}_n(g), \beta) := \min\left\{M : \text{Prob}(n_0 + n \leq M) \geq 1 - \beta\right\}. \tag{13}$$

Here, the cost is assumed to depend on the integrand, g. For the algorithm described in Theorem 4 we may derive an upper bound on the cost by means of Proposition 10 as follows:

$$\begin{aligned}
&\text{cost}(\hat{\mu}_n(g), \beta)\\
&= \min\left\{M : \text{Prob}(n_0 + n \leq M) \geq 1 - \beta\right\}\\
&= n_0 + \min\left\{M : \text{Prob}\left(\min\left(N_C(\epsilon, \alpha_1, \hat{\sigma}^2_{\text{up}}), N_B(\epsilon, \alpha_1, \hat{\sigma}^2_{\text{up}}, \kappa^{3/4}_{\max})\right) \leq M\right) \geq 1 - \beta\right\}.
\end{aligned}$$

From form Proposition 10 it follows that

$$\begin{aligned}
1 - \beta &\leq \text{Prob}\left[\hat{v}_n < \sigma^2\left\{1 + \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\beta}{\beta n}\right)}\right\}\right]\\
&= \text{Prob}\left[\frac{\hat{v}_n}{1 - \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha_1}{\alpha_1 n}\right)}} < \sigma^2\left\{\frac{1 + \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\beta}{\beta n}\right)}}{1 - \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha_1}{\alpha_1 n}\right)}}\right\}\right]\\
&= \text{Prob}\left[\hat{\sigma}^2_{\text{up}}(\hat{v}_{n_0}, n_0, \alpha_1, \kappa_{\max}) < \sigma^2 \gamma(n_0, \alpha_1, \beta, \kappa_{\max})\right],
\end{aligned}$$

where,

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

Since $N_C(\epsilon, \alpha_1, \cdot)$ and $N_B(\epsilon, \alpha_1, \cdot, \kappa^{3/4}_{\max})$ are increasing functions, it follows that

$$\begin{aligned}
\text{cost}(\hat{\mu}_n(g), \beta) \leq n_0 + \\
\min\left(N_C(\epsilon, \alpha_1, \text{var}(g)\gamma(n_0, \alpha_1, \beta, \kappa_{\max})), N_B(\epsilon, \alpha_1, \text{var}(g)\gamma(n_0, \alpha_1, \beta, \kappa_{\max}), \kappa^{3/4}_{\max})\right).
\end{aligned} \tag{14}$$

Theorem 5. *The algorithm described in Theorem 4 has a probabilistic cost given by* (14)*.*

The key factors that determine $\text{cost}(\hat{\mu}_n(g), \beta)$ are ϵ, the error tolerance, and $\text{var}(g)$. The cost is roughly proportional to $\text{var}(g)\epsilon^{-2}$. For the set of integrands $\mathcal{G}_{4,\kappa_{\max}}$ the variance, $\text{var}(g)$ 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.

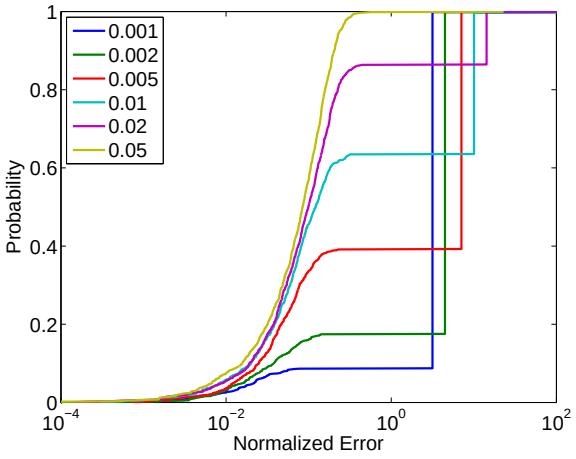

FIGURE 2. Empirical distribution function of $|\mu - \hat{\mu}_n| / \epsilon$ for example (15) with $\mu = \sigma = 1$, $n_0 = 100$, $\kappa_{\max} = 3.2$, $\epsilon = 0.01$, and $p = 0.001, 0.002, 0.005, 0.01, 0.02, 0.05$ using the algorithm in Theorem 4.

3. Example

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

$$g(x) = \begin{cases} 1 + \sigma\sqrt{\frac{1-p}{p}}, & 0 \le x \le p, \\ 1 - \sigma\sqrt{\frac{p}{1-p}}, & p < x \le 1, \end{cases} \tag{15}$$

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

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

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

Figure 2 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.

4. Questions

Here are some questions that suggest themselves:

TABLE 1. Kurtosis 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
$\text{Prob}(\lvert\mu - \hat{\mu}_n\rvert \le \epsilon)$	8.70%	17.50%	39.20%	63.50%	86.40%	99.90%

- 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

Theorem 6 (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*

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

Proof. To prove Chebyshev's inequality note that

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

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

Theorem 7 (Berry-Esseen Inequality)**.** *Let $Y_1, \ldots, Y_n$ be i.i.d. random variables. Suppose that*

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

Then

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

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 8. *Let $\hat{v}_n$ be the sample variance as defined in* (10)*. It's variance is*

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

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

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

$$\hat{v}_n = \frac{1}{n-1}\sum_{i=1}^{n}\left[Y_i - \left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right)\right]^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} Y_i^2 - \sum_{j,k=1}^{n} Y_j Y_k\right]$$

$$\hat{v}_n^2 = \frac{1}{n^2(n-1)^2}\left[n^2\sum_{i,j=1}^{n} Y_i^2 Y_j^2 - 2n\sum_{i,j,k=1}^{n} Y_i^2 Y_j Y_k + \sum_{i,j,k,l=1}^{n} Y_i Y_j Y_k Y_l\right]$$

$$E[Y_i^2 Y_j^2] = \begin{cases} M_4, & i = j, \\ \sigma^4, & i \neq j, \end{cases}$$

$$\sum_{i,j=1}^{n} E[Y_i^2 Y_j^2] = nM_4 + n(n-1)\sigma^4,$$

$$E[Y_i^2 Y_j Y_k] = \begin{cases} M_4, & i = j = k, \\ \sigma^4, & i \neq j, j = k, \\ 0, & j \neq k, \end{cases}$$

$$\sum_{i,j,k=1}^{n} E[Y_i^2 Y_j Y_k] = nM_4 + n(n-1)\sigma^4$$

$$E[Y_i Y_j Y_k Y_l] = \begin{cases} M_4, & i = j = k = l, \\ \sigma^4, & i, j, k, l \text{ have 2 distinct values}, \\ 0, & \text{otherwise}, \end{cases}$$

$$\sum_{i,j,k,l=1}^{n} E[Y_i Y_j Y_k Y_l] = nM_4 + 3n(n-1)\sigma^4$$

$$\begin{aligned} E[\hat{v}_n^2] &= \frac{n^3[M_4 + (n-1)\sigma^4] - 2n^2[M_4 + (n-1)\sigma^4] + n[M_4 + 3(n-1)\sigma^4]}{n^2(n-1)^2} \\ &= \frac{(n-1)M_4 + (n^2 - 2n + 3)\sigma^4}{n(n-1)} \end{aligned}$$

$$\begin{aligned} \operatorname{var}(\hat{v}_n^2) &= E[\hat{v}_n^2] - [E(\hat{v}_n)]^2 = \frac{(n-1)M_4 + (n^2 - 2n + 3)\sigma^4}{n(n-1)} - \sigma^4 \\ &= \frac{(n-1)M_4 + (-n+3)\sigma^4}{n(n-1)} = \frac{1}{n}\left(M_4 - \frac{n-3}{n-1}\sigma^4\right) = \frac{\sigma^4}{n}\left(\kappa - \frac{n-3}{n-1}\right). \end{aligned}$$

□

Theorem 9 (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:*

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

Proof. Define the random variable

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

From conditional probability it is known that

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

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

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

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

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

Combining these results together yields

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

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

Proposition 10. *Let $\hat{v}_n$ be the sample variance of a function g as defined in* (10)*, and let $\kappa = \text{kurt}(g)$. Then*

$$\text{Prob}\left[\hat{v}_n < \sigma^2\left\{1 + \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right\}\right] \geq 1 - \alpha, \tag{16a}$$

$$\text{Prob}\left[\hat{v}_n > \sigma^2\left\{1 - \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right\}\right] \geq 1 - \alpha. \tag{16b}$$

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 9) that

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

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

Then (16a) follows directly. By a similar argument.

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

Thus, (16b) follows as well. □