Monte Carlo Algorithms Where the Integrand Size Is Unknown *

Fred J. Hickernell[1], Lan Jiang[1], Yuewei Liu[2], and Art Owen[3]

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 Introduction

When one tries to utilize theoretical error bounds or complexity results for Monte Carlo integration to inform a practical algorithm, there are some real challenges. To determine the sample size needed, one must have some information about the size of the function, specifically its standard deviation in the case of simple Monte Carlo. Estimating this standard deviation is a step whose error is typically not analyzed. Moreover, if one wants to guarantee that the answer obtained by the Monte Carlo algorithm satisfies a given error tolerance with a high probability, then in practice one often resorts to Central Limit Theorem, whose application also introduces an additional error that is not well understood.

This article attempts to address these shortcomings by derving a tight theory that accounts rigorously for as much of the error in a practical algorithm as is possible.

Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL, USA, hickernell@iit.edu, ljiang14@hawk.iit.edu???other emails · School of Mathematics and Statistics, Lanzhou University, Lanzhou City, Gansu, China 730000, ??? ·

* The first author were partially supported by the National Science Foundation under DMS-1115392

This is done by means of some probability inequalities that involve higher moments and a variant to traditional information-based complexity theory that measures the cost of the problem in terms of the unknown size of the integrand.

2 Simple (I.I.D.) Monte Carlo in Practice

2.1 The Multivariate Integration Problem and a Practical Solution

Suppose one wishes to compute the following integral or mean, μ, of some d-variate 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 specified probability density function. A simple Monte Carlo algorithm to estimate this integral is the sample mean of the function evaluated at independent and identically distributed random variables $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ with the marginal probability density function ρ:

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

A natural question is the following: *How large should n be to ensure with some reasonable certainty, that the absolute error of the approximation is no larger than some positive error tolerance, ε?* That is, for a given a significance level or uncertainty, α how large should n be to make

$$\mathrm{Prob}\,[|\hat{\mu}_n - \mu| \le \varepsilon] \ge 1 - \alpha. \tag{2}$$

In practice, a common approach is to invoke the Central Limit Theorem, which says that

$$\mathrm{Prob}\left[|\hat{\mu}_n - \mu| \le \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right] \approx 1 - \alpha, \tag{3}$$

where z_α denotes the $(1-\alpha)100\%$ percentile of the standard Gaussian distribution, and σ^2 denotes the variance of the function:

$$\sigma^2 := \mathrm{var}(Y_1) = \mathrm{var}(f(\boldsymbol{X}_1)) =: \mathrm{var}(f) = \int_{\mathbb{R}^d} |f(\boldsymbol{x}) - \mu|^2 \rho(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{4}$$

This variance σ^2 is typically unknown a priori, but it can be estimated in terms of the sample variance based on a sample of size, $n_\sigma \in \mathbb{N}$:

$$\hat{\sigma}^2 = \mathfrak{C}^2 \hat{v}_{n_\sigma}, \quad \text{where} \quad \hat{v}_{n_\sigma} = \frac{1}{n_\sigma - 1} \sum_{i=1}^{n_\sigma} [Y_i - \hat{\mu}_{n_\sigma}]^2, \quad \hat{\mu}_{n_\sigma} = \frac{1}{n_\sigma} \sum_{i=1}^{n_\sigma} Y_i, \quad Y_i = f(\boldsymbol{X}_i). \tag{5}$$

The variance inflation factor $\mathfrak{C} > 1$ accounts for the fact that as an unbiased estimator, the sample variance may be larger or smaller than the true variance. Defining

$$N_G(\varepsilon, \alpha) := \left\lceil \left(\frac{z_{\alpha/2}}{\varepsilon} \right)^2 \right\rceil, \tag{6}$$

in light of the Central Limit Theorem, one then estimates the true mean by $\hat{\mu}_n$, based on an additional $n = N_G(\varepsilon/\hat{\sigma}, \alpha)$ independent samples:

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=n_\sigma+1}^{n_\sigma+n} Y_i, \qquad Y_i = f(\boldsymbol{X}_i). \tag{7}$$

This algorithm has a cost of $n_\sigma + N_G(\varepsilon/\hat{\sigma}, \alpha)$ function evaluations, which depends on the initial sample size, n_σ, the error tolerance, ε, the degree of uncertainty, α, the variance inflation factor, $\mathfrak{C}$, and the sample variance, $\hat{\sigma}^2$.

2.2 Lack of Theoretical Justification

The algorithm defined by (5), (6), and (7) is practical. It is economical because it is adaptive, i.e., it estimates the final sample size necessary to approximate the mean. Unfortunately, this algorithm is based on two assumptions that require justification or modification:

i) the variance of $Y = f(\boldsymbol{X})$ is only known approximately, and
ii) the Gaussian approximation for the sample mean given by Central Limit Theorem is only approximately true.

This article proposes to tighten these assumptions. At the same time a complexity theory is developed that allows the cost of computing the answer to be represented in terms of the unknown size of the function, in this case σ.

2.3 Illustrative Univariate Integration Examples

To illustrate where Monte Carlo methods might fail, consider the following univariate step-function integrated against the uniform probability distribution on $[0,1]$, i.e., $\rho = 1$:

$$f(x) = \begin{cases} \mu + \sigma\sqrt{\frac{1-p}{p}}, & 0 \le x \le p, \\ \mu - \sigma\sqrt{\frac{p}{1-p}}, & p < x \le 1, \end{cases} \qquad \mu = \int_0^1 f(x)\,\mathrm{d}x, \qquad \sigma^2 = \int_0^1 [f(x)-\mu]^2\,\mathrm{d}x. \tag{8}$$

where $p \in (0,1)$ is a parameter. An adaptive Monte Carlo algorithm developed in the next section is able to approximate the integral well for moderate values of p but not for very small ones. The test function parameters are $\mu = \sigma = 1$. The algorithm parameters are an absolute error tolerance of 0.01, an uncertainty of $\alpha = 5\%$, a sample size for estimating the variance of $n_\sigma = 1000$, and a variance inflation factor $\mathfrak{C} = 1.5$. Table 1 shows the percentage of times that the error tolerance is met for various values of p. For $p = 0.005$ the algorithm exceeds the required 95% success, but for $p = 0.0001$ the algorithm only gets the correct answer less than 10% of the time. Figure 1 shows the empirical distribution of the normalized error. A normalized error no greater than one means that the error tolerance has been met.

Table 1 Probability of meeting the error tolerance for test function (8) using the adaptive algorithm in Theorem 1.

p	0.0001	0.0002	0.0005	0.001	0.002	0.005
$\mathrm{Prob}(\lvert\mu-\hat{\mu}_n\rvert \le 0.01)$	8.90%	21.30%	39.80%	63.20%	85.80%	99.50%

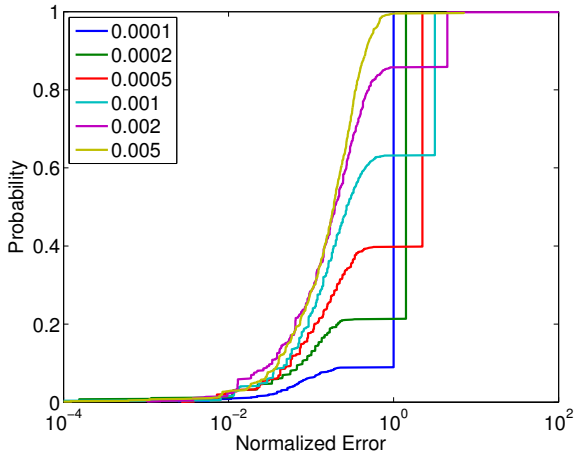

Fig. 1 Empirical distribution function of $|\mu - \hat{\mu}_n|/0.01$ for example (8) and various values of p using the adaptive algorithm in Theorem 1.

The difficulty here is not a deficiency of the adaptive algorithm developed here, which is a modification of what is described above. The point is that any algorithm can be fooled. In this case, the initial sample used to estimate the variance of the

integrand may miss a very narrow spike. Without a good estimate of this variance, there is no reliable determination of the sample size needed for computing a sample mean that is close enough to the mean of the function. Theorem 1 describes under what conditions this adaptive algorithm will not be fooled.

The difficulties of error estimation are not unique to Monte Carlo methods. For the step function f defined in (8), MATLAB's `quad` function approximates $\int_0^1 f(x - 1/\sqrt{2} \pmod 1)\,\mathrm{d}x$ with an error tolerance of 10^{-14} to be ≈ 0.92911, instead of the true answer of 1. Thus, the step function can fool automatic quadrature routines. Figure 2 displays the integrand

$$f(x) = 1 + \cos\left(8\pi \min\left(\max\left(\frac{x - 0.27158}{0.45684}, 0\right), 1\right)\right), \qquad \int_0^1 f(x)\,\mathrm{d}x = 1.54316.$$

Here the constants 0.27158 and 0.45684 are chosen to fool MATLAB's `quad` function. Applying `quad` to approximate the above integral with an error tolerance of 10^{-14} gives the answer 2, since all of the integrand values sampled by `quad` are 2. Thus, `quad` is fooled again, even though this integrand has continuous first derivative.

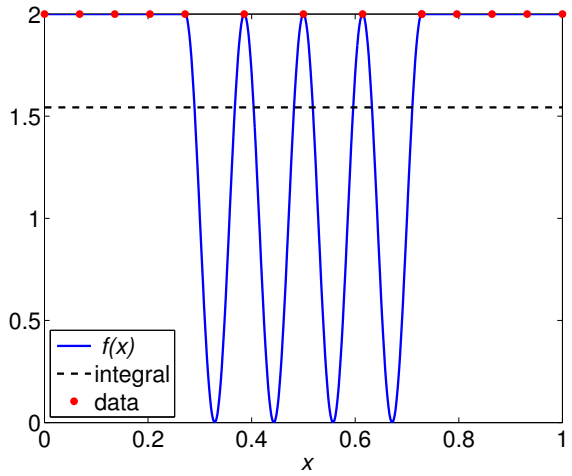

Fig. 2 Integrand that fools MATLAB's `quad` function.

These two examples illustrate that although quadrature rules are valuable, they may give the wrong answers. The goal of theory is to develop quadrature rules that have iron-clad sufficient conditions for the success. This article does so for adaptive Monte Calo rules. Thus, the step function example will be shown to fail to satisfy those conditions when the adaptive Monte Carlo rule fails to meet the required tolerance.

3 Simple (I.I.D.) Monte Carlo with Guaranteed Error Estimation

3.1 Satisfying the Error Tolerance

First the question of bounding the variance is addressed. The integrands under consideration are assumed to have finite moments of up to a order, p, i.e., a finite $\mathscr{L}_p$ norm defined as follows:

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

The absolute centered p^{th} moment of f is defined as

$$M_p(f) := \int_{\mathbb{R}^d} |f(\boldsymbol{x}) - \mu|^p \rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \|f - \mu\|_p^p, \qquad p \geq 1. \tag{9}$$

By these definitions, the variance of Y_1 is the second moment, i.e., $M_2(f) = \mathrm{var}(Y_1) = \sigma^2$, and all functions in $\mathscr{L}_2$ have finite variance. Note that if $1 \leq q \leq p$, then by Hölder's inequality,

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

Thus, the $\mathscr{L}_p$ norms and M_p moments are related as follows:

$$\|f\|_q \leq \|f\|_p, \quad M_q(f) \leq [M_p(f)]^{q/p}, \quad \mathscr{L}_p \subseteq \mathscr{L}_q \qquad \text{for } 1 \leq q \leq p. \tag{10}$$

As mentioned in (5), a practical upper bound on this variance is obtained from the sample variance, $\hat{\sigma}^2 = \mathfrak{C}^2 \hat{v}_{n_\sigma}$, where $\mathfrak{C} > 1$ is some variance inflation factor. This can be justified in a probabilistic sense by appealing to Cantelli's inequality (Theorem 7) and the variance of $\hat{v}_{n_\sigma}$ given by Theorem 6. Proposition 8 implies that

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

where κ is the *kurtosis* of the integrand, defined as

$$\kappa := \mathrm{kurt}(f) = \frac{M_4(f)}{\mathrm{var}^2(f)} = \frac{M_4(f)}{\sigma^4} = \frac{M_4(f)}{M_2^2(f)} \geq 1 \qquad \forall f \in \mathscr{L}_4. \tag{11}$$

Note that the kurtosis of a function is independent of its scale, i.e., $\mathrm{kurt}(cf) = \mathrm{kurt}(f)$ for any $c \neq 0$. Moreover, functions in $\mathscr{L}_4$ automatically have finite variance. It follows that the kurtosis of the integrand must be small enough, relative to n_σ and $\mathfrak{C}$ to ensure that the variance estimate is correct in a probabilistic sense, namely,

$$\frac{1}{1-\sqrt{\left(\kappa-\frac{n_\sigma-3}{n_\sigma-1}\right)\left(\frac{1-\alpha}{\alpha n_\sigma}\right)}} \leq \mathfrak{C}^2$$

$$\iff \kappa \leq \frac{n_\sigma-3}{n_\sigma-1}+\left(\frac{\alpha n_\sigma}{1-\alpha}\right)\left(1-\frac{1}{\mathfrak{C}^2}\right)^2 =: \kappa_{\max}(\alpha, n_\sigma, \mathfrak{C}). \quad (12)$$

Figure 3a shows how large a kurtosis can be accommodated for a given n_σ, α, and variance inflation factor $\mathfrak{C} = 1.5$. Note that for $n = 30$, a common rule of thumb for applying the central limit theorem, even $\alpha = 0.1$ gives $\kappa_{\max}$ of only about 2, which is rather restrictive.

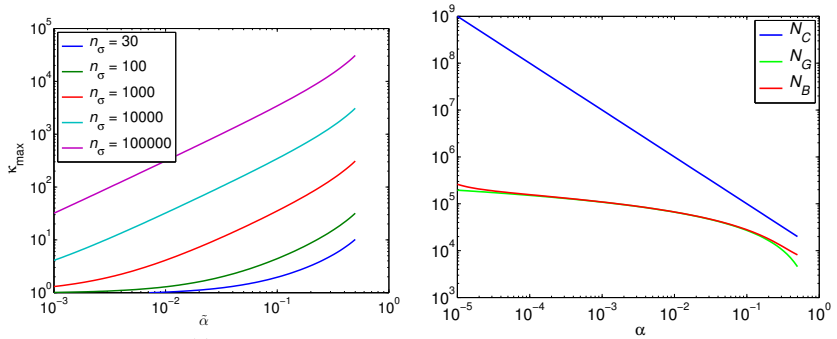

Fig. 3 (a) The maximum kurtosis, $\kappa_{\max}(\alpha, n_\sigma, 1.5)$, as defined in (12); (b) comparison of sample sizes $N_G(0.01, \alpha)$, $N_C(0.01, \alpha)$, and $N_B(0.01, \alpha, \kappa_{\max}^{3/4}(\alpha, 1000, 1.5))$.

The other issue that needs to be addressed is a tight probabilistic error bound. The error bound given by the Central Limit Theorem, (3), is only approximate. Chebyshev's inequality implies that the number of function evaluations needed to ensure that $\hat{\mu}_n$ satisfies the error tolerance with high probability is

$$\mathrm{Prob}\left[|\hat{\mu}_n - \mu| \leq \varepsilon\right] \geq 1-\alpha \quad \text{for } n \geq N_C(\varepsilon/\sigma, \alpha),\ f \in \mathscr{L}_2, \quad (13a)$$

where

$$N_C(\varepsilon, \alpha) := \left\lceil \frac{1}{\alpha \varepsilon^2} \right\rceil. \quad (13b)$$

However, this sample size is much larger than that given by the Central Limit Theorem, as shown in Figure 3b.

Since higher order moments of the integrand are required to guarantee an upper bound on the true variance in terms of the sample variance, it is sensible to use these higher order moments to obtain smaller sample sizes. A smaller sample size than (13b) with a rigorous probabilistic bound can be found by invoking the non-uniform Berry-Esseen inequality (Theorem 5). This inequality makes strong assumptions on the distribution of $f(\boldsymbol{X})$, namely, a finite third moment, $M_3 < \infty$. Recalling that $Y_i = f(\boldsymbol{X}_i)$, $\mu = E(Y_i)$, and $\hat{\mu}_n = (Y_1 + \cdots + Y_n)/n$, it then follows by the non-uniform Berry-Esseen inequality, that

$$\begin{aligned}
\text{Prob}\left[|\hat{\mu}_n - \mu| < \frac{\sigma}{\sqrt{n}}x\right] &= \text{Prob}\left[\hat{\mu}_n - \mu < \frac{\sigma}{\sqrt{n}}x\right] - \text{Prob}\left[\hat{\mu}_n - \mu < -\frac{\sigma}{\sqrt{n}}x\right] \\
&\geq \left[\Phi(x) - \frac{0.56\widetilde{M}_3}{\sqrt{n}}(1+|x|)^{-3}\right] - \left[\Phi(-x) + \frac{0.56\widetilde{M}_3}{\sqrt{n}}(1+|x|)^{-3}\right] \\
&= 1 - 2\left(\frac{0.56\widetilde{M}_3}{\sqrt{n}}(1+|x|)^{-3} + \Phi(-x)\right), \qquad (14)
\end{aligned}$$

where $\widetilde{M}_3 := M_3/\sigma^3$, and Φ is the standard Gaussian cumulative distribution function. Letting $x = \varepsilon\sqrt{n}/\sigma$, the probability of making an error less than ε is bounded below by $1-\alpha$, i.e.,

$$\text{Prob}[|\hat{\mu}_n - \mu| < \varepsilon] \geq 1 - \alpha, \quad \text{provided } n \geq N_B(\varepsilon/\sigma, \alpha, \widetilde{M}_3),\ f \in \mathscr{L}_3, \qquad (15a)$$

where

$$N_B(b, \alpha, M) := \min\left\{ n \in \mathbb{N} : \Phi\left(-b\sqrt{n}\right) + \frac{0.56M}{\sqrt{n}\left(1 + b\sqrt{n}\right)^3} \leq \frac{\alpha}{2} \right\}. \qquad (15b)$$

As shown in Figure 3b, N_B is quite close to N_G for a range of α, but N_B may be somewhat larger for very small or rather α. In general N_B is smaller than N_C, but not always. A disadvantage of (15) is that class of integrands, $\mathscr{L}_3$, is smaller than that in (13), but this typically a small price to pay given the much smaller cost of computation.

The theorem below combines the results on estimating the variance with the sample sizes arising from Chebyshev's inequality and the Berry-Esseen inequality. These lead to an adaptive Monte Carlo algorithm with a probabilistic error guarantee.

Theorem 1. *Specify the following parameters defining the algorithm:*

- *sample size for variance estimation,* $n_\sigma \in \mathbb{N}$,
- *variance inflation factor for variance estimation,* $\mathfrak{C} \in (1, \infty)$,
- *uncertainty tolerance,* $\alpha \in (0,1)$, *and* $\tilde{\alpha} = 1 - \sqrt{1-\alpha}$, *and*
- *relative error tolerance,* $\varepsilon \in (0,1)$.

Define the set of functions with bounded kurtosis:

$$\mathscr{F}^{\text{kurt}} = \{f \in \mathscr{L}_4 : \text{kurt}(f) = \kappa \leq \kappa_{\max}(n_\sigma, \tilde{\alpha}, \mathfrak{C})\},$$

where $\kappa_{\max}$ *is defined in* (12). *For any* $f \in \mathscr{F}^{\text{kurt}}$, *compute the sample variance,* $\hat{v}_{n_\sigma}$ *using a simple random sample of size* n_σ. *Use this to approximate the variance of* f *by* $\hat{\sigma}^2 = \mathfrak{C}^2 \hat{v}_{n_\sigma}$ *as in* (4). *Next define a sample size*

$$n = N_{CB}(\varepsilon/\hat{\sigma}, \tilde{\alpha}, \kappa_{\max}^{3/4}),$$

where

$$N_{CB}(b, \alpha, M) := \min\left(N_C(b, \alpha), N_B(b, \alpha, M)\right), \tag{16}$$

N_C *is defined in* (13b) *and* N_B *is defined in* (15b). *Compute* $\hat{\mu}_n$, *the simple Monte Carlo estimator of* μ *based on* n *samples, as in* (7). *A probabilistic absolute error bound is given by*

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

Proof. Note that (10) implies that the third moment can be bounded in terms of the fourth moment, namely $\tilde{M}_3 \leq \kappa^{3/4}$. There are three primary random variables in this algorithm: the estimated upper bound on the standard deviation, $\hat{\sigma}$, the sample size to estimate the mean, n, which is an explicit function of $\hat{\sigma}$, and the estimated mean, $\hat{\mu}_n$. By (13) and (15) it then follows that $\text{Prob}\left[|\hat{\mu}_n - \mu| \leq \varepsilon\right] \geq 1 - \tilde{\alpha}$, provided that $\hat{\sigma} \geq \sigma$. Thus,

$$\begin{aligned}
\text{Prob}\left[|\hat{\mu}_n - \mu| \leq \varepsilon\right] &= E_{\hat{\sigma}}\left\{\text{Prob}\left[|\hat{\mu}_n - \mu| \leq \varepsilon \mid \hat{\sigma}\right]\right\} \\
&\geq E_{\hat{\sigma}}\left\{(1 - \tilde{\alpha}) 1_{[\sigma,\infty)}(\hat{\sigma})\right\} \\
&\geq (1 - \tilde{\alpha})(1 - \tilde{\alpha}) = 1 - \alpha,
\end{aligned}$$

since $\hat{\sigma} \geq \sigma$ with probability $1 - \tilde{\alpha}$ by (12). □

3.2 Cost of the Algorithm

The sample size of the adaptive algorithm defined in Theorem 1 is a random variable, and so the cost of this algorithm might best be defined defined probabilistically. Moreover, the cost depends strongly on the σ as well as the ε, and its definition should reflect this dependence.

Let A be any random algorithm defined for a set of integrands $\mathscr{F}$ that takes as its input an error tolerance, ε, an uncertainty level, α, and a procedure for computing values of $f \in \mathscr{F}$. The algorithm then computes an approximation to the integral, $A(f, \varepsilon, \alpha)$, satisfying a probabilistic error criterion,

$$\text{Prob}[|\mu - A(f, \varepsilon, \alpha)| < \varepsilon] \leq 1 - \alpha,$$

The integral approximation, $A(f,\varepsilon,\alpha)$ is based solely on $f(\boldsymbol{x}_1),\dots,f(\boldsymbol{x}_n)$, where the choice of each $\boldsymbol{x}_i$ may depend iteratively on $(\boldsymbol{x}_1,f(\boldsymbol{x}_1)),\dots,(\boldsymbol{x}_{i-1},f(\boldsymbol{x}_{i-1}))$, and the decision to stop with n data depends on all function data collected up to and including the n^{th}. Then $\text{card}(A,\varepsilon,\alpha,f) := n$, which as noted is a random variable. The probabilistic cost of the algorithm, with uncertainty β, for integrands of variance no greater than σ^2 is defined as

$$\text{cost}(A,\varepsilon,\alpha,\beta,\mathscr{F},\sigma) := \sup_{\substack{f\in\mathscr{F}\\ \text{var}(f)\le\sigma^2}} \min\{M : \text{Prob}[\text{card}(A,\varepsilon,\alpha,f)\le M]\ge 1-\beta\}. \tag{17}$$

The cost of the particular adaptive Monte Carlo algorithm defined in Theorem 1, denoted aMC, for the class of functions $\mathscr{F}^{\text{kurt}}$ is

$$\text{cost}(\text{aMC},\varepsilon,\alpha,\beta,\mathscr{F}^{\text{kurt}},\sigma) := \sup_{\substack{f\in\mathscr{F}^{\text{kurt}}\\ \text{var}(f)\le\sigma^2}} \min\{M : \text{Prob}(n_\sigma+n\le M)\ge 1-\beta\}. \tag{18}$$

Since n_σ is fixed, bounding this cost depends on bounding n, which depends on $\hat{\sigma}$ as given by Theorem 1. Moreover, $\hat{\sigma}$ can be bounded using Proposition 8:

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

where

$$\gamma^2(\tilde{\alpha},\beta,\mathfrak{C}) := \mathfrak{C}^2\left\{1+\sqrt{\left(\frac{\tilde{\alpha}}{1-\tilde{\alpha}}\right)\left(\frac{1-\beta}{\beta}\right)\left(1-\frac{1}{\mathfrak{C}^2}\right)^2}\right\} > 1.$$

Noting that $N_{CB}(\cdot,\alpha,M)$ is a non-increasing function allows one to derive the following upper bound on the cost of the adaptive Monte Carlo algorithm.

Theorem 2. *The adaptive Monte Carlo algorithm described in Theorem 1, denoted* aMC*, has a probabilistic cost bounded above by*

$$\text{cost}(\text{aMC},\varepsilon,\alpha,\beta,\mathscr{F}^{\text{kurt}},\sigma) \le n_\sigma + N_{CB}(\varepsilon/(\sigma\gamma(\tilde{\alpha},\beta,\mathfrak{C})),\tilde{\alpha},\kappa_{\max}^{3/4}).$$

The cost of the adaptive Monte Carlo algorithm aMC is roughly proportional to $\sigma^2\varepsilon^{-2}$. The set $\mathscr{F}^{\text{kurt}}$ contains integrands with arbitrarily large variance, σ^2 and thus with potentially arbitrarily large algorithmic cost. On the other hand, the cost does seem to depend on the integrand variance as expected. The variable cost of the algorithm for integrands in $\mathscr{F}^{\text{kurt}}$ is actually an advantage, rather than a drawback,

of this analysis. One need not make any a priori assumptions about the size of the integrand, σ, only about the kurtosis, κ, which is unchanged when the integrand is multiplied by an arbitrary nonzero constant.

Whereas σ can be thought of as the size of the integrand, the kurtosis is a measure of the integrands *nastiness*. The larger the kurtosis, the more difficult it is to estimate σ, and thus choose the proper sample size to obtain the desired error tolerance.

4 Simple Monte Carlo Satisfying a Relative Error Criterion

In many practical situations, one needs to approximate the integral with a certain relative accuracy. For example, one wants an answer that is correct to three significant digits. In this case, given a tolerance, ε, and a significance level α, with $\varepsilon, \alpha \in (0,1)$, one seeks a random $\tilde{\mu}$ such that

$$\text{Prob}\left[\left|\frac{\tilde{\mu}-\mu}{\mu}\right| \leq \varepsilon\right] \geq 1-\alpha. \tag{19}$$

Clearly, one must have $\mu \neq 0$ for such a statement to be possible. Using straightforward algebraic manipulations, this condition may be written equivalently as

$$\text{Prob}\left[\frac{\tilde{\mu}}{1+\varepsilon\,\text{sign}(\mu)} \leq \mu \leq \frac{\tilde{\mu}}{1-\varepsilon\,\text{sign}(\mu)}\right] \geq 1-\alpha. \tag{20}$$

The above form is not a traditional confidence interval, however, suppose that one has the following confidence interval for μ in terms of $\hat{\mu}$ and $\tilde{\varepsilon}$:

$$1-\alpha \leq \text{Prob}\left[|\hat{\mu}-\mu| \leq \tilde{\varepsilon}\right] = \text{Prob}\left[\hat{\mu}-\tilde{\varepsilon} \leq \mu \leq \hat{\mu}+\tilde{\varepsilon}\right], \quad \text{with} \quad 0 < \tilde{\varepsilon}/|\hat{\mu}| \leq \varepsilon. \tag{21}$$

Letting $\tilde{\mu} = \hat{\mu}(1-\tilde{\varepsilon}^2/\hat{\mu}^2)$, it follows that

$$\begin{aligned} 1-\alpha \leq \text{Prob}\left[\hat{\mu}-\tilde{\varepsilon} \leq \mu \leq \hat{\mu}+\tilde{\varepsilon}\right] &= \text{Prob}\left[\frac{\tilde{\mu}}{1+\tilde{\varepsilon}/\hat{\mu}} \leq \mu \leq \frac{\tilde{\mu}}{1-\tilde{\varepsilon}/\hat{\mu}}\right] \\ &\leq \text{Prob}\left[\frac{\tilde{\mu}}{1+\varepsilon\,\text{sign}(\hat{\mu})} \leq \mu \leq \frac{\tilde{\mu}}{1-\varepsilon\,\text{sign}(\hat{\mu})}\right] \end{aligned}$$

Since (21) implies that $\text{sign}(\mu) = \text{sign}(\hat{\mu})$, the relative error criterion (20), and thus (19), are satisfied. The previous section shows how to find absolute error criteria of the form (21), but the challenge is to ensure that $\tilde{\varepsilon} \leq \varepsilon\,|\hat{\mu}|$ when $\hat{\mu}$ is not known in advance. This is done iteratively as described in Theorem 3 below.

Some notation is needed for this theorem. For any fixed $\alpha \in (0,1)$, and $M > 0$, define the inverse of the functions $N_C(\cdot,\alpha)$, $N_B(\cdot,\alpha,M)$, and $N_{CB}(\cdot,\alpha,M)$,

$$N_C^{-1}(n,\alpha) := \frac{1}{\sqrt{n\alpha}}, \tag{22a}$$

$$N_B^{-1}(n,\alpha,M) := \min\left\{ b>0 : \Phi\left(-b\sqrt{n}\right) + \frac{0.56M}{\sqrt{n}\left(1+b\sqrt{n}\right)^3} \le \frac{\alpha}{2} \right\}, \tag{22b}$$

$$N_{CB}^{-1}(n,\alpha,M) := \min(N_C^{-1}(n,\alpha), N_B^{-1}(n,\alpha,M)). \tag{22c}$$

It then follows then by Chebyshev's inequality and the Berry-Esseen Inequality (see Theorem 5 and (14)) that

$$\text{Prob}[|\hat{\mu}_n - \mu| < \tilde{\varepsilon}] \ge 1-\alpha, \quad \text{provided } f \in \mathscr{L}_3, \qquad \text{where } \tilde{\varepsilon} = \sigma N_{CB}^{-1}(n,\alpha,\widetilde{M}_3), \tag{22d}$$

and $\widetilde{M}_3$ is the scaled absolute third moment of the integrand. Given a significance level, $\alpha \in (0,1)$, let $\alpha_\sigma, \alpha_1, \alpha_2, \ldots$ be an infinite sequence of positive numbers such that

$$(1-\alpha_\sigma)(1-\alpha_1)(1-\alpha_2)\cdots = 1-\alpha. \tag{23}$$

Fore example, one might choose

$$\alpha_\sigma = 1-\mathrm{e}^{-b}, \;\; \alpha_i = 1-\mathrm{e}^{-ba^{-i}}, \; i \in \mathbb{N}, \quad \text{where } a \in (1,\infty), \; b = \frac{1-a}{a}\log(1-\alpha). \tag{24}$$

Theorem 3. *Specify the following parameters defining the algorithm:*

- *sample size for variance estimation,* $n_\sigma \in \mathbb{N}$,
- *initial sample size for mean estimation,* $n_1 \in \mathbb{N}$,
- *variance inflation factor for variance estimation,* $\mathfrak{C} \in (1,\infty)$,
- *factor for confidence interval width reduction,* $\delta \in (0,1)$,
- *uncertainty tolerance,* $\alpha \in (0,1)$, *and a sequence* $\alpha_\sigma, \alpha_1, \alpha_2, \ldots$ *satisfying* (23), *and*
- *relative error tolerance,* $\varepsilon \in (0,1)$.

Define the set of functions with bounded kurtosis and nonzero mean:

$$\mathscr{F}_0^{\text{kurt}} = \{f \in \mathscr{L}_4 : \text{kurt}(f) = \kappa \le \kappa_{\max}(n_\sigma, \alpha_\sigma, \mathfrak{C}), \mu(f) \ne 0\},$$

where $\kappa_{\max}$ *is defined in* (12). *For any* $f \in \mathscr{F}_0^{\text{kurt}}$, *compute the sample variance,* $\hat{v}_{n_\sigma}$ *using a simple random sample of size* n_σ. *Use this to approximate the variance of* f *by* $\hat{\sigma}^2 = \mathfrak{C}^2\hat{v}_{n_\sigma}$, *as in* (4), *and to compute the width of initial the confidence interval for the mean,* $\tilde{\varepsilon}_1 = \hat{\sigma}N_{CB}^{-1}(n_1,\alpha_1,\kappa_{\max}^{3/4})$. *For* $i = 1,2,\ldots$, *do the following:*

a) Compute the sample average $\hat{\mu}_{n_i}$ *using a simple random sample that is independent of those used to compute* $\hat{v}_{n_\sigma}$ *and* $\hat{\mu}_{n_1}, \ldots, \hat{\mu}_{n_{i-1}}$.

b) If $\tilde{\varepsilon}_i > \varepsilon|\hat{\mu}_{n_i}|$, *then reduce the width of the next confidence interval for the mean,*

$$\tilde{\varepsilon}_{i+1} = \min(\delta\tilde{\varepsilon}_i, \varepsilon\max(\tilde{\varepsilon}_i, |\hat{\mu}_{n_i}| - \tilde{\varepsilon}_i)).$$

Define the next sample size, $n_{i+1} = N_{CB}(\tilde{\varepsilon}_{i+1}/\hat{\sigma}, \alpha_{i+1}, \kappa_{\max}^{3/4})$*, increase* i *by one, and go to step a).*

c) Else, let $\tilde{\mu} = \hat{\mu}_{n_i}(1 - \tilde{\varepsilon}_i^2/\hat{\mu}_{n_i}^2)$*, and terminate the algorithm because the relative error criterion,* (19)*, is satisfied.*

Proof. In this algorithm there are a number of important random variables: the estimated upper bound on the standard deviation, $\hat{\sigma}$, the sample sizes $n_1, \ldots, n_\tau$, the number of confidence intervals computed, τ, and the estimates of the mean, $\hat{\mu}_{n_1}, \ldots, \hat{\mu}_{n_\tau}$. These sample means are conditionally independent given the sequence of sample sizes. The probability that the final confidence interval is correct, is then no less than the probability that all of the confidence intervals are correct, conditioned on the sample sizes. Specifically,

$$\begin{aligned}
\text{Prob}\left[\left|\frac{\tilde{\mu}-\mu}{\mu}\right| \le \varepsilon\right] &\ge \text{Prob}\left[|\hat{\mu}_{n_\tau} - \mu| \le \tilde{\varepsilon}_{n_\tau} \;\&\; \varepsilon\hat{\mu}_{n_\tau} \le \tilde{\varepsilon}_{n_\tau}\right] \\
&= E\left\{\text{Prob}\left[|\hat{\mu}_{n_\tau} - \mu| \le \tilde{\varepsilon}_{n_\tau} \;\&\; \varepsilon\hat{\mu}_{n_\tau} \le \tilde{\varepsilon}_{n_\tau} \mid \hat{\sigma}, \tau, n_1, \ldots, n_\tau\right]\right\} \\
&\ge E\left\{\text{Prob}\left[|\hat{\mu}_{n_i} - \mu| \le \tilde{\varepsilon}_{n_i}\right] \forall i \;\&\; \varepsilon\hat{\mu}_{n_\tau} \le \tilde{\varepsilon}_{n_\tau} \mid \hat{\sigma}, \tau, n_1, \ldots, n_\tau\right]\right\} \\
&\ge E_{\hat{\sigma}}\left\{[(1-\alpha_1)(1-\alpha_2)\cdots]1_{[\sigma,\infty)}(\hat{\sigma})\right\} \\
&\ge (1-\alpha_\sigma)(1-\alpha_1)(1-\alpha_2)\cdots = 1-\alpha. \qquad \square
\end{aligned}$$

5 Numerical Examples

6 Questions

Here are some questions that suggest themselves:

- 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.

7 Appendix of Useful Results

Theorem 4 (Chebyshev's Inequality). *Let Z be any random variable with mean μ and variance σ^2. Then for all $\alpha > 0$,*

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

Proof. To prove Chebyshev's inequality note that

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

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

Theorem 5 (Non-uniform Berry-Esseen Inequality). *(Petrov, 1995, Theorem 5.16, p. 168) Let $Y_1,\ldots,Y_n$ be i.i.d. random variables. Suppose that $\mu = E(Y_i)$, $\text{var}(Y_i) = \sigma^2 > 0$, and $\widetilde{M}_3 = E\left|Y_i-\mu\right|^3/\sigma^3 < \infty$. Then*

$$\left|\text{Prob}\left[\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(Y_i-\mu) < x\right] - \Phi(x)\right| \leq \frac{A\widetilde{M}_3}{\sqrt{n}}(1+|x|)^{-3}.$$

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

Theorem 6. *Let $Y_1, Y_2, \ldots$ be i.i.d. random variables, $\hat{\mu}_n$ be the sample mean as defined in* (1)*, and $\hat{v}_n$ be the sample variance as defined in* (5)*. Then the variance of the sample variance and the kurtosis of the sample mean are given by*

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

where κ is the kurtosis of Y_1.

Proof. Without loss of generality, it may be assumed that the Y_i have zero mean. If they do not, the mean may be subtracted off. The sample variance and its square may be written in terms of multiple sums of the Y_i as follows:

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

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

The expected values of each of these multiple sums can be computed in terms of $\sigma^2 = E[Y_i^2]$ and $\kappa = E[Y_i^4]/\sigma^4$:

$$E[Y_i^2 Y_j^2] = \begin{cases} \kappa\sigma^4, & i = j, \\ \sigma^4, & i \neq j, \end{cases} \qquad \sum_{i,j=1}^{n} E[Y_i^2 Y_j^2] = n\sigma^4(\kappa + n - 1),$$

$$E[Y_i^2 Y_j Y_k] = \begin{cases} \kappa\sigma^4, & i = j = k, \\ \sigma^4, & i \neq j, j = k, \\ 0, & j \neq k, \end{cases} \qquad \sum_{i,j,k=1}^{n} E[Y_i^2 Y_j Y_k] = n\sigma^4(\kappa + n - 1)$$

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

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

This last expectations yields the kurtosis of the sample mean:

$$\text{kurt}(\hat{\mu}_n) = \frac{E\left[\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i\right\}^4\right]}{[\text{var}(\hat{\mu}_n)]^2} = \frac{n^{-4} E\left[\sum_{i,j,k,l=1}^{n} Y_i Y_j Y_k Y_l\right]}{\sigma^4/n^2} = \frac{\kappa + 3(n-1)}{n}$$

Combining these expectations also yields the expected value for $\hat{v}_n^2$ and the variance of $\hat{v}_n$, since the sample variance has mean σ^2/n:

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

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

□

Theorem 7 (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

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

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 8 *Let $\hat{v}_n$ be the sample variance of a function g as defined in* (5)*, and let $\kappa = \operatorname{kurt}(g)$. Then*

$$\operatorname{Prob}\left[\hat{v}_n < \sigma^2\left\{1+\sqrt{\left(\kappa-\frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right\}\right] \geq 1-\alpha, \tag{25a}$$

$$\operatorname{Prob}\left[\hat{v}_n > \sigma^2\left\{1-\sqrt{\left(\kappa-\frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right\}\right] \geq 1-\alpha. \tag{25b}$$

Proof. Choosing

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

it follows from Cantelli's inequality (Theorem 7) that

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

Then (25a) follows directly. By a similar argument, applying Cantelli's inequality to the expression Prob $\left[-\hat{v}_n + \sigma^2 \geq a\right]$ implies (25b). □

Proposition 9 *For the situation described in Section* **??** *with the operations* $\oplus$, $\ominus$, *and* $\otimes$, *the basis functions* $\phi_{\boldsymbol{k}}$, *and any* $\oplus$*-closed set, P, with cardinality* $|P|$ *and associated dual lattice* $P^{\perp}$, *it follows that*

$$\mu(\phi_{\boldsymbol{k}}) = \int_{[0,1)^d} \phi_{\boldsymbol{k}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{[0,1)^d} e^{2\pi\sqrt{-1}\boldsymbol{k}\otimes\boldsymbol{x}}\,\mathrm{d}\boldsymbol{x} = \delta_{\boldsymbol{k},\boldsymbol{0}}, \tag{26}$$

$$Q(\phi_{\boldsymbol{k}}, P, \boldsymbol{\Delta}) = \frac{1}{|P|}\sum_{\boldsymbol{x}\in P} \phi_{\boldsymbol{k}}(\boldsymbol{x}\oplus\boldsymbol{\Delta}) = \begin{cases} \phi_{\boldsymbol{k}}(\boldsymbol{\Delta}), & \boldsymbol{k}\in P^{\perp} \\ 0, & \boldsymbol{k}\notin P^{\perp} \end{cases} \tag{27}$$

Proof. The properties assumed for the binary operations imply that for all $\boldsymbol{t} \in [0,1)^d$,

$$\int_{[0,1)^d} \phi_{\boldsymbol{k}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{[0,1)^d} \phi_{\boldsymbol{k}}(\boldsymbol{x}\oplus\boldsymbol{t})\,\mathrm{d}\boldsymbol{x} = \phi_{\boldsymbol{k}}(\boldsymbol{t})\int_{[0,1)^d} \phi_{\boldsymbol{k}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

$$0 = [\phi_{\boldsymbol{k}}(\boldsymbol{t}) - 1]\int_{[0,1)^d} \phi_{\boldsymbol{k}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

This inequality is satisfied iff $\boldsymbol{k} = \boldsymbol{0}$ or $\int_{[0,1)^d} \phi_{\boldsymbol{k}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0$. This establishes (26).

A similar argument establishes (27). Note first that the sum can be simplified to a multiple of $\phi_{\boldsymbol{k}}(\boldsymbol{\Delta})$:

$$\frac{1}{|P|}\sum_{\boldsymbol{x}\in P} \phi_{\boldsymbol{k}}(\boldsymbol{x}\oplus\boldsymbol{\Delta}) = \frac{1}{|P|}\sum_{\boldsymbol{x}\in P} [\phi_{\boldsymbol{k}}(\boldsymbol{x})\phi_{\boldsymbol{k}}(\boldsymbol{\Delta})] = \phi_{\boldsymbol{k}}(\boldsymbol{\Delta})\frac{1}{|P|}\sum_{\boldsymbol{x}\in P} \phi_{\boldsymbol{k}}(\boldsymbol{x}).$$

Next, note that for all $\boldsymbol{t} \in P$ the remaining sum above is unchanged when the argument of $\phi_{\boldsymbol{k}}$ is shifted by $\boldsymbol{t}$:

$$\frac{1}{|P|}\sum_{\boldsymbol{x}\in P} \phi_{\boldsymbol{k}}(\boldsymbol{x}) = \frac{1}{|P|}\sum_{\boldsymbol{x}\in P} \phi_{\boldsymbol{k}}(\boldsymbol{x}+\boldsymbol{t}) = \frac{1}{|P|}\sum_{\boldsymbol{x}\in P} [\phi_{\boldsymbol{k}}(\boldsymbol{x})\phi_{\boldsymbol{k}}(\boldsymbol{t})] = \phi_{\boldsymbol{k}}(\boldsymbol{t})\frac{1}{|P|}\sum_{\boldsymbol{x}\in P} \phi_{\boldsymbol{k}}(\boldsymbol{x}),$$

$$0 = (\phi_{\boldsymbol{k}}(\boldsymbol{t}) - 1)\,\frac{1}{|P|}\sum_{\boldsymbol{x}\in P} \phi_{\boldsymbol{k}}(\boldsymbol{x}).$$

This inequality is satisfied iff $\phi_{\boldsymbol{k}}(\boldsymbol{t}) = 0$ for all $\boldsymbol{t} \in P$, i.e., $\boldsymbol{k} \in P^{\perp}$, or $\sum_{\boldsymbol{x}\in P} \phi_{\boldsymbol{k}}(\boldsymbol{x}) = 0$. This implies (27). □

References

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

8 Quasi-Standard Error Estimation for Quasi-Monte Carlo Methods

Quasi-Monte Carlo methods utilize low discrepancy node sets, in particular integration lattices and digital nets, to expedite the convergence of the sample mean to the true mean the sample size increases. Like the case of i.i.d. sampling discussed in the previous section, one is faced with the question of how to choose the sample size in a data-driven way.

One way used in practice is to compute independent randomizations of the low discrepancy sequence. This means replacing $Y_i = f(\boldsymbol{X}_i)$ in the equation for the sample mean, (7), and the equation for the variance estimate, (5), which are used to define the adaptive algorithm of Theorem 1, by

$$Y_i = \frac{1}{n_1} \sum_{j=1}^{n_1} f(\boldsymbol{Z}_{ij}).$$

Here the set $\{\boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{in_1}\}$ is the i^{th} independent randomization (replication) of an integration lattice or net with n_1 points.

If n_1 is fixed, then Theorems 1 and 2 may be extended in a natural way. Now the number of integrand values required is n_1 times what the formulas in those theorems state, however, the total cost in practice may well be less than simple Monte Carlo because the sample variance using n_σ replications of a low discrepancy node set with n_1 points may be much smaller than the corresponding sample variance using i.i.d. sampling with $n_\sigma n_1$ integrand values. However, for this approach of fixed n_1, the sample variance only tells you how to choose *more replications*. It cannot tell you how to increase the size of the low discrepancy set while keeping the number of replications fixed. Thus, the benefits of quasi-Monte Carlo methods are not fully realized.

If the size of the low discrepancy set is increased until the error tolerance is met, then the algorithm becomes somewhat more complex. Let, $n_1, n_2, \ldots$ be a sequence of sample sizes for which one may obtain low discrepancy node sets. Moreover, let α_{11},

Here there is an additional challenge.

9 Quasi-Standard Error Estimation for Quasi-Monte Carlo Methods

An alternative is to use the *quasi-standard error*, also called *internal replications* to estimate the error and then determine a reasonable sample size. The quasi-standard error has been proposed by **?**, but its implementation and justification have been criticized by **?**. A closer look at the quasi-standard error specifically applied to lattices and nets is provided here.

Integration lattices **references** A general framework that encompasses both kinds of designs is employed so that the analysis of both can be combined. Table 2 provides the definitions for the general notation introduced below. For integration lattices the operations are done with modulo arithmetic, and for digital nets they are done with b-ary digit-wise arithmetic.

Table 2 Definitions for low discrepancy designs for $x,t \in [0,1)$, $\boldsymbol{x},\boldsymbol{t} \in [0,1)^d$, $k,l \in \Omega^d$ and $\boldsymbol{k},\boldsymbol{l} \in \Omega^d$

Integration Lattices	Digital Nets in Base b
$x \oplus t := x + t \bmod 1$	$x \oplus t = \sum_{\ell=1}^{\infty} x_\ell b^{-\ell} \oplus \sum_{\ell=1}^{\infty} t_\ell b^{-\ell} := \sum_{\ell=1}^{\infty} (x_\ell + t_\ell \bmod b) b^{-\ell} \pmod 1$
$\ominus x := 1 - x \bmod 1$	$\ominus \sum_{\ell=1}^{\infty} x_\ell b^{-\ell} := \sum_{\ell=1}^{\infty} (b - x_\ell \bmod b) b^{-\ell} \pmod 1$
$\boldsymbol{x} \oplus \boldsymbol{t} = (x_1,\dots,x_d) \oplus (t_1,\dots,t_d) = (x_1 \oplus t_1,\dots,x_d \oplus t_d)$	
$\ominus \boldsymbol{x} = \ominus(x_1,\dots,x_d) = (\ominus x_1,\dots,\ominus x_d)$	
$\Omega := \mathbb{Z}$	$\Omega := \mathbb{N}_0$
$k \oplus l := k + l$	$k \oplus l = \sum_{\ell=0}^{\infty} k_\ell b^\ell \oplus \sum_{\ell=0}^{\infty} l_\ell b^\ell := \sum_{\ell=0}^{\infty} (k_\ell + l_\ell \bmod b) b^\ell$
$\ominus k := -k$	$\ominus k = \ominus \sum_{\ell=0}^{\infty} k_\ell b^\ell := \sum_{\ell=0}^{\infty} (b - k_\ell \bmod b) b^\ell$
$\boldsymbol{k} \oplus \boldsymbol{l} = (k_1,\dots,k_d) \oplus (l_1,\dots,l_d) = (k_1 \oplus l_1,\dots,k_d \oplus l_d)$	
$\ominus \boldsymbol{k} = \ominus(k_1,\dots,k_d) = (\ominus k_1,\dots,\ominus k_d)$	
$k \otimes x = kx \bmod 1$	$k \otimes x = \sum_{\ell=0}^{\infty} k_\ell b^\ell \otimes \sum_{\ell=1}^{\infty} x_\ell b^{-\ell} := \sum_{\ell=0}^{\infty} k_\ell x_{\ell+1} \pmod 1$
$\boldsymbol{k} \otimes \boldsymbol{x} = (k_1,\dots,k_d) \otimes (x_1,\dots,x_d) = k_1 \otimes x_1 + \cdots + k_d \otimes x_d \pmod 1$	

Consider the problem of integration over the half-open unit cube with respect to the uniform density, i.e., $\rho = 1_{[0,1)^d}$. Define a binary operation $\oplus : [0,1)^d \times [0,1)^d \to [0,1)^d$ that is commutative, i.e., $\boldsymbol{x} \oplus \boldsymbol{t} = \boldsymbol{t} \oplus \boldsymbol{x}$ for all $\boldsymbol{x},\boldsymbol{t} \in [0,1)^d$. It is not necessarily the case that $\oplus$ is associative. For example, for the case corresponding to digital nets in base 2,

$$(1/3 \oplus 2/3) \oplus 2/3 = ({}_2 0.\overline{01} \oplus {}_2 0.\overline{10}) \oplus {}_2 0.\overline{10} = ({}_2 0.\overline{11} \bmod 1) \oplus {}_2 0.\overline{10} = 0 \oplus {}_2 0.\overline{10} = 2/3,$$

whereas

$$1/3 \oplus (2/3 \oplus 2/3) = {}_2 0.\overline{01} \oplus ({}_2 0.\overline{10} \oplus {}_2 0.\overline{10}) = {}_2 0.\overline{01} \oplus {}_2 0.\overline{00} = {}_2 0.\overline{01} = 1/3.$$

However, it is assumed that for some set $\widetilde{[0,1)}^d \subseteq [0,1)^d$, associativity does hold, i.e.,

$$(\boldsymbol{x} \oplus \boldsymbol{t}) \oplus \boldsymbol{y} = \boldsymbol{x} \oplus (\boldsymbol{t} \oplus \boldsymbol{y}) \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in [0,1)^d, \ \boldsymbol{t} \in \widetilde{[0,1)}^d.$$

Assume that shifting a measurable set $B \subseteq [0,1)^d$ by $\boldsymbol{x}$, keeps its volume unchanged, i.e., $\mu(1_{B \oplus \boldsymbol{x}}) = \mu(1_B)$. Also define $\ominus : [0,1)^d \to [0,1)^d$ such that $\boldsymbol{x} \oplus (\ominus \boldsymbol{x}) = \boldsymbol{0}$ for all $\boldsymbol{x} \in [0,1)^d$, and let $\boldsymbol{x} \ominus \boldsymbol{t}$ be a short-cut for $\boldsymbol{x} \oplus (\ominus \boldsymbol{t})$. A $\oplus$-closed node set, $P \subset \widetilde{[0,1)}^d$ is a finite cardinality group under $\oplus$. For any such node set, P, and a shift $\boldsymbol{\Delta} \in [0,1)^d$, a shifted cubature rule is defined as

$$Q(f; P, \boldsymbol{\Delta}) = \frac{1}{|P|} \sum_{\boldsymbol{x} \in P} f(\boldsymbol{x} \oplus \boldsymbol{\Delta}),$$

where $|P|$ denotes the cardinality of P.

The proposed integration rule using lattices or net node sets is based on an infinite sequences of embedded $\oplus$-closed node sets,

$$P_0 \subset P_1 \subset P_2 \subset \cdots, \tag{28}$$

and an offset, m. Since these node sets are groups, it follows that they can be written as direct sums, $P_{\ell+m} = P_\ell \oplus \widetilde{P}_{\ell m}$, where the $\widetilde{P}_{\ell m}$ are $\oplus$-closed node sets of cardinality $|P_{\ell+m}| \,/\, |P_\ell|$. The quasi-standard error is defined analogously to the standard error for i.i.d. sampling:

$$\operatorname{qse}(f; P_{\ell+m}, \boldsymbol{\Delta}, m) = \sqrt{\frac{1}{|\widetilde{P}_{\ell m}|(|\widetilde{P}_{\ell m}| - 1)} \sum_{\boldsymbol{t} \in \widetilde{P}_{\ell m}} [Q(f; P_\ell, \boldsymbol{t} \oplus \boldsymbol{\Delta}) - Q(f; P_{\ell+m}, \boldsymbol{\Delta})]^2}. \tag{29}$$

Given an error tolerance, ε, and a variance inflation factor, $\mathfrak{C}$, the algorithm computes the approximation to the integral $Q(f; P_{\ell+m}, \boldsymbol{\Delta})$ and the error estimate $\mathfrak{C} \cdot qse(f; P_{\ell+m}, \boldsymbol{\Delta}, m)$ for $\ell = 0, 1, \ldots$ until the error estimate is no greater than the tolerance, i.e.,

$$\hat{\mu} = Q(f; P_n, \boldsymbol{\Delta}), \qquad \text{where } n = \min\{\ell + m : \mathfrak{C} \cdot \operatorname{qse}(f; P_{\ell+m}, \boldsymbol{\Delta}, m) \le \varepsilon\}. \tag{30}$$

The argument must now be made to show for what kinds of integrands, this algorithm works and how much it will cost. To this end the space of possible integrands, $\mathscr{F}$, is defined as the vector space of functions that can be expressed as absolutely convergent sequences:

$$f = \sum_{\boldsymbol{k} \in \Omega^d} \hat{f}_{\boldsymbol{k}} \phi_{\boldsymbol{k}}, \qquad \left\| (\hat{f}_{\boldsymbol{k}})_{\boldsymbol{k} \in \Omega^d} \right\|_1 < \infty, \qquad \phi_{\boldsymbol{k}}(\boldsymbol{x}) = \mathrm{e}^{2\pi\sqrt{-1}\boldsymbol{k} \otimes \boldsymbol{x}} \quad \forall \boldsymbol{x} \in [0,1)^d.$$

Here $\Omega \subseteq \mathbb{Z}$ is the set of wavenumbers, and $\otimes : \Omega^d \times [0,1)^d \to [0,1)$ is like a dot product. It is assumed to satisfy

$$\boldsymbol{k}\otimes\boldsymbol{x}=0 \iff \boldsymbol{k}=\boldsymbol{0} \text{ or } \boldsymbol{x}=\boldsymbol{0}, \qquad \forall \boldsymbol{k}\in\Omega^d,\ \boldsymbol{x}\in[0,1)^d,$$

$$\boldsymbol{k}\otimes(\boldsymbol{x}\oplus\boldsymbol{t})=(\boldsymbol{k}\otimes\boldsymbol{x})+(\boldsymbol{k}\otimes\boldsymbol{t}) \bmod 1 \qquad \forall \boldsymbol{k}\in\Omega^d,\ \boldsymbol{x}\in\widetilde{[0,1)}^d,\ \boldsymbol{t}\in[0,1)^d.$$

Note that the space $\mathscr{F}$ is shift invariant, i.e., $f\in\mathscr{F} \implies f(\cdot\oplus\boldsymbol{t})\in\mathscr{F}$ for all $\boldsymbol{t}\in[0,1)^d$. Moreover, the notation $\oplus$ and $\ominus$ are extended to Ω^d such that

$$\boldsymbol{k}\oplus\boldsymbol{l}=\boldsymbol{l}\oplus\boldsymbol{k}, \quad \boldsymbol{k}\oplus(\ominus\boldsymbol{k})=\boldsymbol{0}, \quad (\boldsymbol{k}\oplus\boldsymbol{l})\otimes\boldsymbol{x}=(\boldsymbol{k}\otimes\boldsymbol{x})+(\boldsymbol{l}\otimes\boldsymbol{x}) \bmod 1 \quad \forall \boldsymbol{k},\boldsymbol{l}\in\Omega^d,\ \boldsymbol{x}\in[0,1)^d.$$

The properties assumed here for these binary operations imply that

$$\phi_{\boldsymbol{0}}=1, \quad \phi_{\boldsymbol{k}}\phi_{\boldsymbol{l}}=\phi_{\boldsymbol{k}\oplus\boldsymbol{l}}, \quad \phi_{\boldsymbol{k}}(\boldsymbol{x}\oplus\boldsymbol{t})=\phi_{\boldsymbol{k}}(\boldsymbol{x})\phi_{\boldsymbol{k}}(\boldsymbol{t}), \qquad \forall \boldsymbol{k},\boldsymbol{l}\in\Omega^d,\ \boldsymbol{x}\in\widetilde{[0,1)}^d,\ \boldsymbol{t}\in[0,1)^d.$$

As shown in Proposition 9 they also imply that $\int_{[0,1)^d}\phi_{\boldsymbol{k}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=\delta_{\boldsymbol{k},\boldsymbol{0}}$. These two facts combine to yield the orthogonality of the $\phi_{\boldsymbol{k}}$ and an integral expression for the series coefficients:

$$\int_{[0,1)^d}\phi_{\boldsymbol{k}}(\boldsymbol{x})\phi_{\ominus\boldsymbol{l}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=\delta_{\boldsymbol{k}\ominus\boldsymbol{l},\boldsymbol{0}}, \qquad \hat{f}_{\boldsymbol{k}}=\int_{[0,1)^d}f(\boldsymbol{x})\phi_{\ominus\boldsymbol{k}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

For any $\oplus$-closed node set P, the corresponding dual set of wavenumbers, $P^\perp$, figures prominently in the error analysis of $Q(f;P,\boldsymbol{\Delta})$. The dual set is defined as

$$P^\perp:=\{\boldsymbol{k}\in\Omega^d:\boldsymbol{k}\otimes\boldsymbol{x}=0\ \forall\boldsymbol{x}\in P\}.$$

Note that by definition, $\boldsymbol{0}\in P^\perp$. The dual set minus the zero vector is denoted $P^{\perp\prime}$. Proposition 9 then allows one to express the integral of a function, the approximation to the integral, and the error, all conveniently in terms of the series coefficients:

$$\begin{aligned}
\mu(f)&=\sum_{\boldsymbol{k}\in\Omega_d}\hat{f}_{\boldsymbol{k}}\mu(\phi_{\boldsymbol{k}})=\hat{f}_{\boldsymbol{0}}=(S_{\boldsymbol{0}}f)(\boldsymbol{\Delta})\\
Q(f;P,\boldsymbol{\Delta})&=\sum_{\boldsymbol{k}\in\Omega_d}\hat{f}_{\boldsymbol{k}}Q(\phi_{\boldsymbol{k}};P,\boldsymbol{\Delta})=\sum_{\boldsymbol{k}\in P^\perp}\hat{f}_{\boldsymbol{k}}\phi_{\boldsymbol{k}}(\boldsymbol{\Delta})=(S_{P^\perp}f)(\boldsymbol{\Delta})\\
e(f;P,\boldsymbol{\Delta})&:=\mu(f)-Q(f;P,\boldsymbol{\Delta})=-\sum_{\boldsymbol{k}\in P^{\perp\prime}}\hat{f}_{\boldsymbol{k}}\phi_{\boldsymbol{k}}(\boldsymbol{\Delta})=-(S_{P^{\perp\prime}}f)(\boldsymbol{\Delta}),
\end{aligned} \tag{31}$$

where for any $\Xi\subseteq\Omega^d$, the filtering operator $S_\Xi:\mathscr{F}\to\mathscr{F}$ keeps only the terms with wavenumbers contained in Ξ:

$$S_\Xi f:=\sum_{\boldsymbol{k}\in\Xi}\hat{f}_{\boldsymbol{k}}\phi_{\boldsymbol{k}}.$$

The quasi-standard error can also be expressed in terms of the dual set and the filtering operator. From the definition in (29) it follows that

$$\begin{aligned}
\mathrm{qse}^2(f;P_{\ell+m},\boldsymbol{\Delta},m) &= \frac{1}{|\widetilde{P}_{\ell m}|(|\widetilde{P}_{\ell m}|-1)} \sum_{\boldsymbol{t}\in\widetilde{P}_{\ell m}} [e(f;P_\ell,\boldsymbol{t}\oplus\boldsymbol{\Delta})-e(f;P_{\ell+m},\boldsymbol{\Delta})]^2 \\
&= \frac{1}{|\widetilde{P}_{\ell m}|-1}\left[\frac{1}{|\widetilde{P}_{\ell m}|}\sum_{\boldsymbol{t}\in\widetilde{P}_{\ell m}} e^2(f;P_\ell,\boldsymbol{t}\oplus\boldsymbol{\Delta})-e^2(f;P_{\ell+m},\boldsymbol{\Delta})\right] \\
&= \frac{1}{|\widetilde{P}_{\ell m}|-1}\left[\frac{1}{|\widetilde{P}_{\ell m}|}\sum_{\boldsymbol{t}\in\widetilde{P}_{\ell m}} (S_{P_\ell^{\perp\prime}}f)^2(\boldsymbol{t}\oplus\boldsymbol{\Delta})-(S_{P_{\ell+m}^{\perp\prime}}f)^2(\boldsymbol{\Delta})\right].
\end{aligned}$$

Since $\widetilde{P}_{\ell m}$ is a $\oplus$-closed set, the sum above can be rewritten as a filtering of the square of the filtered integrand evaluated at the shift:

$$\frac{1}{|\widetilde{P}_{\ell m}|}\sum_{\boldsymbol{t}\in\widetilde{P}_{\ell m}} (S_{P_\ell^{\perp\prime}}f)^2(\boldsymbol{t}\oplus\boldsymbol{\Delta}) = Q((S_{P_\ell^{\perp\prime}}f)^2,\widetilde{P}_{\ell m},\boldsymbol{\Delta}) = S_{\widetilde{P}_{\ell m}^\perp}\left((S_{P_\ell^{\perp\prime}}f)^2\right)(\boldsymbol{\Delta}).$$

Since the integrand is real-valued, the square of its filtered version may be written as

$$\begin{aligned}
(S_{P_\ell^{\perp\prime}}f)^2 &= \sum_{\boldsymbol{k},\boldsymbol{l}\in P_\ell^{\perp\prime}} \hat{f}_{\boldsymbol{k}}\hat{f}_{\boldsymbol{l}}^*\phi_{\boldsymbol{k}}\phi_{\ominus\boldsymbol{l}} = \sum_{\boldsymbol{k},\boldsymbol{l}\in P_\ell^{\perp\prime}} \hat{f}_{\boldsymbol{k}}\hat{f}_{\boldsymbol{l}}^*\phi_{\boldsymbol{k}\ominus\boldsymbol{l}} \\
S_{\widetilde{P}_{\ell m}^\perp}\left((S_{P_\ell^{\perp\prime}}f)^2\right) &= \sum_{\substack{\boldsymbol{k},\boldsymbol{l}\in P_\ell^{\perp\prime}\\ \boldsymbol{k}\ominus\boldsymbol{l}\in\widetilde{P}_{\ell m}^\perp}} \hat{f}_{\boldsymbol{k}}\hat{f}_{\boldsymbol{l}}^*\phi_{\boldsymbol{k}\ominus\boldsymbol{l}} = \sum_{\substack{\boldsymbol{k},\boldsymbol{l}\in P_\ell^{\perp\prime}\\ \boldsymbol{k}\ominus\boldsymbol{l}\in P_{\ell+m}^\perp}} \hat{f}_{\boldsymbol{k}}\hat{f}_{\boldsymbol{l}}^*\phi_{\boldsymbol{k}\ominus\boldsymbol{l}}
\end{aligned}$$

since $\boldsymbol{k},\boldsymbol{l}\in P_\ell^{\perp\prime}$ implies $\boldsymbol{k}\ominus\boldsymbol{l}\in P_\ell^\perp$ and $P_\ell^\perp\cap\widetilde{P}_{\ell m}^\perp = P_{\ell+m}^\perp$. The dual set $P_\ell^\perp$ may be written as a direct sum, $P_{\ell+m}^\perp\oplus\hat{P}_{\ell m}$, for some $\hat{P}_{\ell m}\subset\Omega^d$ with the same cardinality as $\widetilde{P}_{\ell m}$. For any pair $\boldsymbol{k},\boldsymbol{l}\in P_\ell^{\perp\prime}$ with $\boldsymbol{k}\ominus\boldsymbol{l}\in P_{\ell+m}^\perp$, it follows that $\boldsymbol{k},\boldsymbol{l}\in P_{\ell+m}^\perp\oplus\{\boldsymbol{r}\}$ for some $\boldsymbol{r}\in\hat{P}_{\ell m}$. This observation allows the sum above to be understood as a sum of squares of filtered versions of the integrand:

$$\begin{aligned}
S_{\widetilde{P}_{\ell m}^\perp}\left((S_{P_\ell^{\perp\prime}}f)^2\right) &= \sum_{\boldsymbol{k},\boldsymbol{l}\in P_{\ell+m}^{\perp\prime}} \hat{f}_{\boldsymbol{k}}\hat{f}_{\boldsymbol{l}}^*\phi_{\boldsymbol{k}\ominus\boldsymbol{l}} + \sum_{\boldsymbol{r}\in\hat{P}_{\ell m}'}\sum_{\boldsymbol{k},\boldsymbol{l}\in P_{\ell+m}^\perp\oplus\{\boldsymbol{r}\}} \hat{f}_{\boldsymbol{k}}\hat{f}_{\boldsymbol{l}}^*\phi_{\boldsymbol{k}\ominus\boldsymbol{l}} \\
&= (S_{P_{\ell+m}^{\perp\prime}}f)^2 + \sum_{\boldsymbol{r}\in\hat{P}_{\ell m}'} (S_{P_{\ell+m}^\perp\oplus\{\boldsymbol{r}\}}f)^2,
\end{aligned}$$

where $\hat{P}_{\ell m}' = \hat{P}_{\ell m}\setminus\{\boldsymbol{0}\}$. This expression then leads to the following expression for the quasi-standard error.

Lemma 1. *For cubature rules based on a sequence of embedded $\oplus$-closed node sets, as in* (28)*, the quasi-standard error as defined in* (29) *may be expressed in terms of filtered versions of the integrand as follows:*

$$\mathrm{qse}(f;P_{\ell+m},\boldsymbol{\Delta},m)=\sqrt{\frac{1}{|\hat{P}_{\ell m}|-1}\sum_{\boldsymbol{r}\in\hat{P}'_{\ell m}}(S_{P^{\perp}_{\ell+m}\oplus\{\boldsymbol{r}\}}f)^2(\boldsymbol{\Delta})}.$$

The absolute error of the cubature rule based on node set $P_{\ell+m}$ with shift $\boldsymbol{\Delta}$ is $|(S_{P^{\perp\prime}_{\ell+m}}f)(\boldsymbol{\Delta})|$ according to (31), i.e., the filtered version of the integrand keeping just wavenumbers in $P^{\perp\prime}_{\ell+m}$ and then evaluated at the shift. The quasi-standard error is root mean square of filtered versions of the integrand using shifts of the dual node set, $P^{\perp}_{\ell+m}$.

In practice, one usually does not know, κ, the kurtosis of the integrand. The choice of n_σ, $\mathfrak{C}$, and α imply a $\kappa_{\max}$ which one is willing to accept. However, there is a way to check whether the implicit assumption about the integrand's kurtosis is reasonable. The sample of size n used to estimate the integral as $\hat{\mu}_n$, may also be used to compute the sample variance, $\hat{v}_n$, which is independent of the sample variance, $\hat{v}_{n_\sigma}$, used to determine the sample size n. Using Cantelli's inequality

$$\begin{aligned}\mathrm{Prob}(\hat{v}_n\geq\hat{\sigma}^2)&=\mathrm{Prob}(\hat{v}_n-L^2\hat{v}_{n_\sigma}\geq 0)\\&=\mathrm{Prob}[\hat{v}_n-L^2\hat{v}_{n_\sigma}-(1-L^2)\sigma^2\geq(L^2-1)\sigma^2]\\&\leq\frac{\mathrm{var}(\hat{v}_n-L^2\hat{v}_{n_\sigma})}{\mathrm{var}(\hat{v}_n-L^2\hat{v}_{n_\sigma})+\{(L^2-1)\sigma^2\}^2}=\frac{1}{1+\frac{(L^2-1)^2\sigma^4}{\mathrm{var}(\hat{v}_n-L^2\hat{v}_{n_\sigma})}}.\end{aligned}$$

This above quotient in the denominator can be further simplified by noticing that $\hat{v}_n$ and $\hat{v}_{n_\sigma}$ are independent. Thus,

$$\begin{aligned}\frac{\mathrm{var}(\hat{v}_n-L^2\hat{v}_{n_\sigma})}{(L^2-1)^2\sigma^4}&=\frac{\mathrm{var}(\hat{v}_n)+L^4\,\mathrm{var}(\hat{v}_{n_\sigma})}{(L^2-1)^2\sigma^4}\\&=\frac{\frac{1}{n}\left(\kappa-\frac{n-3}{n-1}\right)+L^4\frac{1}{n_\sigma}\left(\kappa-\frac{n_\sigma-3}{n_\sigma-1}\right)}{(L^2-1)^2}\\&=\frac{\left(\frac{1}{n}+L^4\frac{1}{n_\sigma}\right)\left(\kappa-\frac{n_\sigma-3}{n_\sigma-1}\right)+\frac{1}{n}\left(\frac{n_\sigma-3}{n_\sigma-1}-\frac{n-3}{n-1}\right)}{(L^2-1)^2}\\&\leq\frac{\left(\frac{1}{n}+L^4\frac{1}{n_\sigma}\right)\left(\frac{\alpha_1 n_\sigma}{1-\alpha_1}\right)\left(1-\frac{1}{L^2}\right)^2-\frac{2}{n}\left(\frac{1}{n_\sigma-1}-\frac{1}{n-1}\right)}{(L^2-1)^2}\\&=\left(1+\frac{n_\sigma}{nL^4}\right)\left(\frac{\alpha_1}{1-\alpha_1}\right)-\frac{2(n-n_\sigma)}{n(n_\sigma-1)(n-1)(L^2-1)^2},\end{aligned}$$

which implies that

$$\begin{aligned}\operatorname{Prob}(\hat{v}_n \geq \hat{\sigma}^2) &\leq \frac{\left(1+\frac{n_\sigma}{nL^4}\right)\left(\frac{\alpha_1}{1-\alpha_1}\right) - \frac{2(n-n_\sigma)}{n(n_\sigma-1)(n-1)(L^2-1)^2}}{1+\left(1+\frac{n_\sigma}{nL^4}\right)\left(\frac{\alpha_1}{1-\alpha_1}\right) - \frac{2(n-n_\sigma)}{n(n_\sigma-1)(n-1)(L^2-1)^2}} \\ &\leq \left(1+\frac{n_\sigma}{nL^4}\right)\left(\frac{\alpha_1}{1-\alpha_1}\right) \quad \text{for } n \geq n_\sigma. \end{aligned} \tag{32}$$

This inequality shows that $\hat{v}_n \geq \hat{\sigma}^2$ with a small probability. Thus, if $\hat{v}_n \geq \hat{\sigma}^2$ occurs in practice, then one may have reason to question whether $\sigma^2 \leq \hat{\sigma}^2$, and thus question the implicit assumption on the kurtosis. Figure 4 shows the upper bound on this probability for typical choices of α_1, L, n_σ, and n.

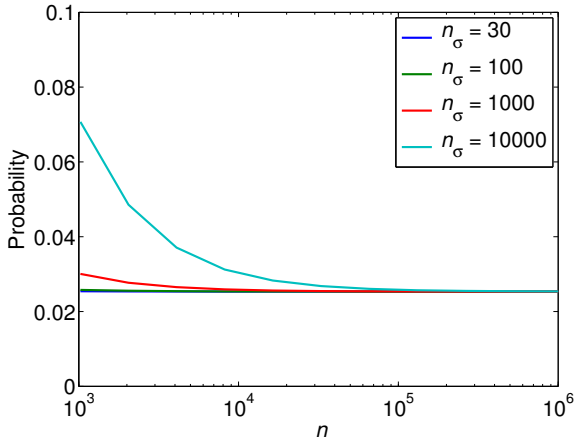

Fig. 4 The upper bound on the probability that $\hat{v}_n \geq \hat{\sigma}^2$ in (32) for $\alpha_1 = 1 - \sqrt{95\%} \approx 2.5\%$ and $L = 1.5$.