Approximate Fixed Width Confidence Intervals Via Monte Carlo Sampling *

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

[1] Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL, USA, hickernell@iit.edu,ljiang14@hawk.iit.edu
[2] School of Mathematics and Statistics, Lanzhou University, Lanzhou City, Gansu, China 730000, ???
[3] Art's address here with email here

Summary. When we are done, we will write this.

1 Introduction

Monte Carlo algorithms provide a flexible way to approximate $\mu = \mathbb{E}(Y)$ when one can generate samples of the random variable Y. For example, Y might be the discounted payoff of some financial derivative, which depends on the future performance of assets that are described by a stochastic model. Then μ is the fair option price. The goal is to obtain a *confidence interval*

$$\Pr[|\mu - \hat{\mu}_n| \le \varepsilon] \ge 1 - \alpha, \tag{1}$$

where

- μ is approximated by
$$\hat{\mu} = \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i, \tag{2}$$
the sample average of n independent and identically distributed (IID) samples of Y,
- ε is the half-width of the confidence interval, which also serves as an *error tolerance*, and
- α is the level of *uncertainty*, e.g., 1% or 0.1%, which is fixed in advance.

Often the sample size, n, is fixed in advance, and the central limit theorem (CLT) provides an approximate value for ε in terms of n and

$$\sigma^2 = \mathrm{Var}(Y) = \mathbb{E}[(Y-\mu)^2], \tag{3}$$

which itself may be approximated by the sample variance.

The goal here is somewhat different. We want to fix ε in advance and then determine how large the sample size must be to obtain a fixed width confidence intervals of the form (1). In this paper we present Algorithm 1 for obtaining such a fixed width confidence interval for the

* The first and second authors were partially supported by the National Science Foundation under DMS-0923111 and DMS-1115392

mean of a real random variable, which is suitable for Monte Carlo sampling. Before presenting the method, we outline the reasons that existing fixed width intervals are not suitable.

The width (equivalently length) of a confidence interval tends to become smaller as the number n of sampled function values increases. In special circumstance, we can choose n to get a confidence interval of at most the desired length and at least the desired coverage level, $1-\alpha$. For instance, if the variance, $\sigma^2 = \mathrm{Var}(Y)$, is known then an approach based on Chebychev's inequality is available, though the actual coverage will usually be much higher than the nominal level, meaning that much narrower intervals would have sufficed. Known variance in addition to a Gaussian distribution for the function values supports a fixed width interval construction that is not too conservative. The CLT provides confidence interval that is asymptotically correct, but our aim is for something that is definitely correct for finite sample sizes. Finally, conservative fixed width intervals for means can be constructed for bounded random variables, by appealing to exponential inequalities Hoeffding's or Chernoff's inequality.

If the relevant variance or bound is unknown, then approaches based on sequential statistics (Siegmund, 1985) may be available. In sequential methods one keeps increasing n until the interval is narrow enough. Sequential confidence intervals require us to take account of the stopping rule when computing the confidence level. They are available in special circumstances, such as Gaussian or binary data. Similarly, Bayesian methods can support a fixed width interval containing μ with $1-\alpha$ posterior probability, and Bayesian methods famously do not require one to account for stopping rules. They do however require strong distributional assumptions.

The solutions described above require strong assumptions that generally do not hold in Monte Carlo applications. The form of the distribution for Y is generally not known, $\mathrm{Var}(Y)$ is generally not known, and Y is not necessarily bounded. There is no assumption-free way to obtain exact confidence intervals for a mean, as has been known since (Bahadur and Savage, 1956, Corollary 2). Some kind of assumption is needed to rule out settings where the desired quantity is the mean of a heavy tailed random variable in which rarely seen large values dominate the mean and spoil the estimates of the variance. The assumption we use is an upper bound on the kurtosis (normalized fourth moment) of the random variable Y:

$$\tilde{\kappa} = \frac{\mathbb{E}[(Y-\mu)^4]}{\sigma^4} \le \tilde{\kappa}_{\max}. \tag{4}$$

Under such an assumption we present a two stage algorithm: the first stage generates a conservative variance estimate, and the second stage uses this variance estimate and a Berry-Esseen Theorem, which can be thought of as a non-asymptotic CLT, to determine how large n must be for the sample mean to satisfy the confidence interval (1). Theorem ??? then demonstrates the validity of the the fixed-width confidence interval, and Theorem ??? demonstrates that the cost of this algorithm is reasonable.

One might question whether assumption (4), which involves fourth-order moments of Y, is more reasonable than an assumption involving only the second moment of Y. For example, using Chebychev's inequality with the assumption

$$\sigma^2 \le \sigma^2_{\max} \tag{5}$$

also yields a fixed-width confidence interval of the form (1). We would argue that (4) is indeed more reasonable. Firstly, if Y satisfies (4), then so does cY for any real constant c, however, the analog does not hold for (5). In fact, if σ is nonzero (5) is violated by cY for c sufficiently large. Second of all, making $\tilde{\kappa}_{\max}$ a factor of 10 or 100 larger than $\tilde{\kappa}$ does not significantly affect the total cost (number of samples required) of our Monte Carlo Algorithm 1 for a large

range of values of σ/ε. However, the cost of our Monte Carlo algorithm, and indeed any Monte Carlo algorithm is proportional to σ^2, so overestimating σ^2 by a factor of 10 or 100 to be safe, increases the cost of the algorithm by that factor.

An important special case of computing $\mu = \mathbb{E}(Y)$ arises in the situation where $Y = f(\boldsymbol{X})$ for some d-variate function $f : \mathbb{R}^d \to \mathbb{R}$, and some d-vector random variable, $\boldsymbol{X}$ with probability density $\rho : \mathbb{R}^d \to [0,\infty)$. One may then interpret the mean of Y as the multidimensional integral

$$\mu = \mathbb{E}(Y) = \mu(f) = \int_{\mathbb{R}^d} f(\boldsymbol{x})\rho(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{6}$$

Given the problem of evaluating $\mu = \int_{\mathbb{R}^d} g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$, one must choose a probability density ρ for which one can easily generate random variates $\boldsymbol{X}$, and then set $f = g/\rho$. The quantities σ^2 and $\tilde{\kappa}$ defined above can be written in terms of weighted $\mathscr{L}_p$-norms of f:

$$\|f\|_p := \left\{ \int_{\mathbb{R}^d} |f(\boldsymbol{x})|^p \rho(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \right\}^{1/p}, \qquad \sigma^2 = \|f-\mu\|_2^2, \qquad \tilde{\kappa} = \frac{\|f-\mu\|_4^4}{\|f-\mu\|_2^4}. \tag{7}$$

For a given g, the choice of ρ is not unique, and making an optimal choice belongs to the realm of *importance sampling*. The assumption of bounded kurtosis, (4), required by Algorithm 1 corresponds to an assumption about the integrand f lying in the *cone* of functions

$$\mathscr{C}_{\tilde{\kappa}_{\max}} = \{ f \in \mathscr{L}_4 : \|f-\mu\|_4 \le \tilde{\kappa}_{\max}^{1/4} \|f-\mu\|_2 \}. \tag{8}$$

From the perspective of numerical analysis, if ρ has independent marginals, one may apply a product form of a univariate quadrature rule to evaluate μ. However, this consumes a geometrically increasing number of samples as d increases, and moreover, such methods often require rather strict smoothness assumptions on f. If f satisfies moderate smoothness conditions, then (randomized) quasi-Monte Carlo methods, or low discrepancy sampling (Niederreiter, 1992; Sloan and Joe, 1994; Lemieux, 2009; Dick and Pillichshammer, 2010) methods for evaluating μ that are more efficient than simple Monte Carlo, however, practical error estimation for these methods remains a challenge.

Computational mathematicians have also addressed the problem of constructing automatic algorithms, i.e., given an error tolerance of ε, one computes an approximation, $\hat{\mu}_n$, based on n evaluations of the integrand f, such that $|\mu - \hat{\mu}_n| \le \varepsilon$. For example, MATLAB (The MathWorks, Inc., 2012), a popular numerical package, contains `quad`, an adaptive Simpson's rule for univariate quadrature routine developed by Gander and Gautschi (2000). Although this and other automatic rules that we are aware work well in practice, they do not have any rigorous guarantees that the error tolerance is met, and it is relatively simple to construct functions that fool them. This is discussed in Section ???. Since a random algorithm, like Monte Carlo, gives a random answer, any statements about satisfying an error criterion must be probabilistic. This leads us back to the problem of finding a fixed-width confidence interval, (1).

An outline of this paper follows.

2 Background probability and statistics

In our Monte Carlo applications, a quantity of interest is written as an expectation: $\mu = \mathbb{E}(Y)$, where Y is a real valued random variable. As mentioned above, very often $Y = f(\boldsymbol{X})$ where $\boldsymbol{X} \in \mathbb{R}^d$ is a random vector with probability density function ρ. In other settings the random

quantity $\boldsymbol{X}$ might have a discrete distribution or be infinite dimensional (e.g,. a Gaussian process) or both. For Monte Carlo estimation, we can work with the distribution of Y alone. The Monte Carlo estimate of μ is the sample mean, as given in (2), where the Y_i are IID random variables with the same distribution as Y.

2.1 Moments

Our methods require conditions on the first four moments of Y as described here. The variance of Y, as defined in (3), is denoted by σ^2, and its non-negative square root, σ, is the standard deviation of Y. Some of our expressions assume without stating it that $\sigma > 0$, and all will require $\sigma < \infty$. The skewness of Y is $\gamma = \mathbb{E}((Y-\mu)^3)/\sigma^3$, and the kurtosis of Y is $\kappa = \tilde{\kappa} - 3 = \mathbb{E}((Y-\mu)^4)/\sigma^4 - 3$ (see (4)). If $\sigma = 0$, then $\gamma = 0$, and $\kappa = -2$ by convention. The mysterious 3 in κ is there to make it zero for Gaussian random variables. Also, $\mu, \sigma^2, \gamma, \kappa$ are related to the first four cumulants (McCullagh, 1987, p.??) of the distribution of Y, meaning that

$$\log(\mathbb{E}[\exp(tY)]) = \mu t + \frac{\sigma^2 t^2}{2} + \frac{\sigma^3 \gamma t^3}{3!} + \frac{\sigma^4 \kappa t^4}{4!} + o(t^4).$$

Our main results require that $\kappa < \infty$, which then implies that *sigma* and γ are finite.

In addition to the moments above we will also use some centered absolute moments of the form $M_k = M_k(Y) = \mathbb{E}(|Y-\mu|^k)$. Normalized versions of these are $\widetilde{M}_k = M_k(Y)/\sigma^k$. In particular, $\widetilde{M}_4 = \tilde{\kappa}$ and $\widetilde{M}_3$ governs the convergence rate of the CLT.

It is a standard result that $1 \le q \le p < \infty$ implies $M_q(Y) \le M_p(Y)^{q/p}$ and similarly $\widetilde{M}_q(Y) \le \widetilde{M}_p(Y)^{q/p}$. The special case

$$\widetilde{M}_3(Y) \le \widetilde{M}_4(Y)^{3/4} \tag{9}$$

will be important for us.

2.2 CLT intervals

A random variable Z has the standard normal distribution, denoted by $\mathcal{N}(0,1)$, if

$$\Pr(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} =: \Phi(z) \exp(-t^2/2)\,\mathrm{d}t.$$

Under the central limit theorem, the distribution of $\sqrt{n}(\hat{\mu}_n - \mu)/\sigma$ approaches $\mathcal{N}(0,1)$ as $n \to \infty$. As a result

$$\Pr\bigl(\hat{\mu}_n - 2.58\sigma\sqrt{n} \le \mu \le \hat{\mu}_n + 2.58\sigma/\sqrt{n}\bigr) \to 0.99 \tag{10}$$

as $n \to \infty$. We write the interval in (10) as $\hat{\mu}_n \pm 2.58\sigma/\sqrt{n}$. Equation (10) is not usable when σ^2 is unknown, but the usual estimate

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \hat{\mu}_n)^2 \tag{11}$$

may be substituted, yielding the interval $\hat{\mu}_n \pm 2.58 s_n/\sqrt{n}$ which also satisfies the limit in (10) by Slutsky's theorem. For an arbitrary confidence level $1-\alpha \in (0,1)$, we replace the constant 2.58 by $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$. The width of this interval is $2z_{\alpha/2}s_n/\sqrt{n}$, and when μ is in the interval then the absolute error $|\mu - \hat{\mu}_n| \le \varepsilon := z_{\alpha/2}s_n/\sqrt{n}$.

Reference for Slutsky's Thm?

I switched to z_α as a tail probability because we are using capital letters for random variables

The coverage level of the CLT interval is only asymptotic. In more detail, Hall (1986) shows that

$$\Pr\big(|\mu-\hat{\mu}_n| \le 2.58s/\sqrt{n}\big) = 0.99 + \frac{1}{n}(A+B\gamma^2+C\kappa) + O\Big(\frac{1}{n^2}\Big) \tag{12}$$

Which equation in Hall's paper? I couldn't find it.

for constants A, B, and C that depend on the desired coverage level (here 99%). Hall's theorem requires only that the random variable Y has sufficiently many finite moments and is not supported solely on a lattice (such as the integers). It is interesting to note that the $O(1/n)$ coverage error in (12) is better than the $O(1/\sqrt{n})$ root mean squared error for the estimate $\hat{\mu}_n$ itself.

2.3 Standard Probability Inequalities

Here we present some well known inequalities that we will make use of. First, Chebychev's inequality ensures that a random variable (such as $\hat{\mu}_n$) is seldom too far from its mean.

Theorem 1 (Chebychev's Inequality). *(Lin and Bai, 2010, 6.1c, p. 52) Let Z be a random variable with mean μ and variance $\sigma^2 \ge 0$. Then for all $\varepsilon > 0$,*

$$\Pr[|Z-\mu| \ge \varepsilon] \le \frac{\sigma^2}{\varepsilon^2}.$$

In some settings we need a one sided inequality like Chebychev's. We will use this one due to Cantelli.

Theorem 2 (Cantelli's Inequality). *(Lin and Bai, 2010, 6.1e, p. 53) Let Z be any random variable with mean μ and finite variance σ^2. For any $a \ge 0$, it follows that:*

$$\Pr[Z-\mu \ge a] \le \frac{\sigma^2}{a^2+\sigma^2}.$$

Berry-Esseen type theorems govern the rate at which a CLT takes hold. We will use the following one.

Theorem 3 (Non-uniform Berry-Esseen Inequality). *(Petrov, 1995, Theorem 5.16, p. 168) Let $Y_1,\dots,Y_n$ be IID random variables with mean μ, variance $\sigma^2 > 0$, and third centered moment $M_3 = E\,|Y_i-\mu|^3/\sigma^3 < \infty$. Then*

$$\left|\Pr\left[\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(Y_i-\mu) < x\right] - \Phi(x)\right| \le \frac{AM_3}{\sqrt{n}(1+|x|)^3} \qquad \forall x \in \mathbb{R},$$

where A is some number satisfying $0.4097 \le A \le 0.5600$.

Our method requires probabilistic bounds on the sample variance, s_n^2. For that, we will use some moments of the variance estimate.

Theorem 4. *(Miller, 1986, ??? and Eq. (7.16)) Let $Y_1,\dots,Y_n$ be IID random variables with variance σ^2 and modified kurtosis $\tilde{\kappa}$ defined in* (4)*. Let s_n^2 be the sample variance as defined in* (11)*. Then the sample variance is unbiased, $\mathbb{E}(s_n^2) = \sigma^2$, and its variance is*

$$\mathrm{Var}(s_n^2) = \frac{\sigma^4}{n}\left(\tilde{\kappa} - \frac{n-3}{n-1}\right).$$

3 Two stage confidence interval

Our two stage procedure works as follows. In the first stage, we take a sample of independent values $Y_1,\ldots,Y_{n_\sigma}$ from the distribution of Y. From this sample we compute the sample variance, $s^2_{n_\sigma}$, according to (11) and estimate the variance of Y_i by $\hat\sigma^2 = \mathfrak{C}^2\hat v_{n_\sigma}$, where $\mathfrak{C}^2 > 1$ is a "variance inflation factor" that will reduce the probability that we have underestimated $\sigma^2 = \mathrm{Var}(Y)$. For the second stage, we use the estimate $\hat\sigma^2$ as if it were the true variance of Y_i and use Berry-Esseen theorem to obtain a suitable sample size, n, for computing the sample average, $\hat\mu_n$, that satisfies the fixed with confidence interval (1).

The next two subsections give details of these two steps that will let us bound their error probabilities. Then we give a theorem on the method as a whole.

3.1 Conservative variance estimates

We need to ensure that our first stage estimate of the variance σ^2 is not too small. The following result bounds the probability of such an underestimate.

Lemma 1. *Let $Y_1,\ldots,Y_n$ be IID random variables with variance $\sigma^2 > 0$ and kurtosis κ. Let s_n^2 be the sample variance defined at* (11)*, and let $\tilde\kappa = \kappa + 3$. Then*

$$\Pr\left[s_n^2 < \sigma^2\left\{1+\sqrt{\left(\tilde\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right\}\right] \ge 1-\alpha, \tag{13a}$$

$$\Pr\left[s_n^2 > \sigma^2\left\{1-\sqrt{\left(\tilde\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right\}\right] \ge 1-\alpha. \tag{13b}$$

Proof. Applying Theorem 4 and choosing

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

it follows from Cantelli's inequality (Theorem 2) that

$$\begin{aligned}\Pr\left[s_n^2 - \sigma^2 \ge \sigma^2\sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right] &= \Pr\left[s_n^2 - \sigma^2 \ge a\right] \\ &\le \frac{\mathrm{Var}(s_n^2)}{a^2 + \mathrm{Var}(s_n^2)} = \frac{\mathrm{Var}(s_n^2)}{\mathrm{Var}(s_n^2)\frac{1-\alpha}{\alpha} + \mathrm{Var}(s_n^2)} = \frac{1}{\left(\frac{1-\alpha}{\alpha}\right)+1} = \alpha.\end{aligned}$$

Then (13a) follows directly. By a similar argument, applying Cantelli's inequality to the expression $\Pr\left[-s_n^2 + \sigma^2 \ge a\right]$ implies (13b). □

Using Lemma 1 we can bound the probability that $\hat\sigma^2 = \mathfrak{C}^2 s^2_{n_\sigma} < \sigma^2$. Equation (13a) implies that

$$\Pr\left[\frac{s^2_{n_\sigma}}{1-\sqrt{\left(\tilde\kappa - \frac{n_\sigma-3}{n_\sigma-1}\right)\left(\frac{1-\alpha}{\alpha n_\sigma}\right)}} > \sigma^2\right] \ge 1-\alpha. \tag{14}$$

Thus, it makes sense for us to require the modified kurtosis, $\tilde{\kappa}$, to be small enough, relative to n_σ, α, and $\mathfrak{C}$, in order to ensure that (14) holds. Specifically, we require

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

or equivalently,

$$\tilde{\kappa} \le \frac{n_\sigma-3}{n_\sigma-1} + \left(\frac{\alpha n_\sigma}{1-\alpha}\right)\left(1-\frac{1}{\mathfrak{C}^2}\right)^2 =: \tilde{\kappa}_{\max}(\alpha, n_\sigma, \mathfrak{C}). \tag{15}$$

This condition is the explicit version of (4) mentioned in the introduction.

3.2 Conservative interval widths

Here we consider how to choose the sample size n to get the desired coverage level from an interval with half-length at most ε. We suppose here that σ is known. In practice we will use a conservative (biased high) estimate for σ.

First, if the CLT held exactly and not just asymptotically, then we could use a CLT sample size, of

$$N_{\mathrm{CLT}}(\varepsilon,\sigma,\alpha) = \left\lceil \left(\frac{z_{\alpha/2}\sigma}{\varepsilon}\right)^2 \right\rceil$$

independent values of Y_i in an interval like the one in (10).

Given knowledge of σ, but no assurance of a Gaussian distribution for $\hat{\mu}_n$, we could instead select a sample size based on Chebychev's inequality (Theorem 1). Taking

$$N_{\mathrm{Cheb}}(\varepsilon,\sigma,\alpha) = \left\lceil \frac{\sigma^2}{\alpha\varepsilon^2} \right\rceil \tag{16}$$

IID observations of Y gives the confidence interval (1). Naturally $N_{\mathrm{Cheb}} \ge N_{\mathrm{CLT}}$.

Finally, we could use the non-uniform Berry-Esseen inequality from Theorem 3. This inequality requires a finite scaled third moment $M_3 = E\left|Y_i-\mu\right|^3/\sigma^3$. The non-uniform Berry-Esseen inequality implies that

$$\begin{aligned}
\Pr\left[|\hat{\mu}_n-\mu| \le \frac{\sigma}{\sqrt{n}}x\right] &= \Pr\left[\hat{\mu}_n-\mu \le \frac{\sigma}{\sqrt{n}}x\right] - \Pr\left[\hat{\mu}_n-\mu < -\frac{\sigma}{\sqrt{n}}x\right] \\
&\ge \left[\Phi(x) - \frac{0.56M_3}{\sqrt{n}(1+|x|)^3}\right] - \left[\Phi(-x) + \frac{0.56M_3}{\sqrt{n}(1+|x|)^3}\right] \\
&= 1 - 2\left(\frac{0.56M_3}{\sqrt{n}(1+|x|)^3} + \Phi(-x)\right),
\end{aligned} \tag{17}$$

Letting $x = \varepsilon\sqrt{n}/\sigma$, the probability of making an error no greater than ε is bounded below by $1-\alpha$, i.e., the fixed width confidence interval (1) holds, provided $n \ge N_B(\varepsilon/\sigma,\alpha,M_3)$, where the Berry-Esseen sample size is

$$N_{\mathrm{BE}}(\varepsilon,\sigma,\alpha,M) := \min\left\{ n \in \mathbb{N} : \Phi\left(-\sqrt{n}\varepsilon/\sigma\right) + \frac{0.56M}{\sqrt{n}\left(1+\sqrt{n}\varepsilon/\sigma\right)^3} \le \frac{\alpha}{2} \right\}. \tag{18}$$

To compute this value, we need to know M_3. In practice, substituting an upper bound on M_3 yields an upper bound on the necessary sample size.

It is possible that in some situations $N_{\mathrm{BE}} > N_{\mathrm{Cheb}}$ might hold. In such cases we could use N_{Cheb} instead.

3.3 Algorithm and Proof of Its Success

In detail, the two stage algorithm works as described below.

Algorithm 1. The user specifies four quantities:

- an initial sample size for variance estimation, $n_\sigma \in \{2,3,\dots\}$,
- a variance inflation factor $\mathfrak{C}^2 \in (1,\infty)$,
- an uncertainty tolerance $\alpha \in (0,1)$, and,
- an error tolerance or confidence interval width, $\varepsilon > 0$.

At the first stage of the algorithm, $Y_1,\dots,Y_{n_\sigma}$ are sampled independently from the same distribution as Y. Then the conservative variance estimate, $\hat\sigma^2 = \mathfrak{C}^2 s_{n_\sigma}^2$ is computed in terms of the sample variance, $s_{n_\sigma}^2$, defined by (11).

To prepare for the second stage of the algorithm we compute $\tilde\alpha = 1-\sqrt{1-\alpha}$ and then $\tilde\kappa_{\max} = \tilde\kappa_{\max}(\tilde\alpha, n_\sigma, \mathfrak{C})$ using equation (15). The sample size for the second stage is

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

where

$$N_\mu(\varepsilon,\sigma,\alpha,M) := \min\bigl(n_\sigma, N_{\text{Cheb}}(\varepsilon,\sigma,\alpha), N_{\text{BE}}(\varepsilon,\sigma,\alpha,M)\bigr). \tag{19}$$

Recall that N_{Cheb} is defined in (16) and N_{BE} is defined in (18).

After this preparation, the second stage is to sample $Y_{n_\sigma+1},\dots,Y_{n_\sigma+n}$ independently from the distribution of Y and compute the sample mean,

$$\hat\mu = \hat\mu_n = \frac{1}{n}\sum_{i=n_\sigma+1}^{n_\sigma+n} Y_i. \tag{20}$$

The success of this algorithm is guaranteed in the following theorem. The main assumption needed is bounded kurtosis.

Theorem 5. *Let Y be a random variable with mean μ and modified kurtosis $\tilde\kappa \le \tilde\kappa_{\max}(\tilde\alpha, n_\sigma, \mathfrak{C})$. It follows that Algorithm 1 above yields an estimate $\hat\mu$ given by* (20) *which satisfies the fixed width confidence interval*

$$\Pr(|\hat\mu - \mu| \le \varepsilon) \ge 1-\alpha.$$

Proof. The first stage yields a variance estimate satisfying $\Pr(\hat\sigma^2 > \sigma^2) \ge 1-\tilde\alpha$ by (15) applied with uncertainty tolerance $\tilde\alpha$. The second stage yields $\Pr(|\hat\mu_n - \mu| \le \varepsilon) \ge 1-\tilde\alpha$ by the Berry-Esseen result (17), so long as $\hat\sigma \ge \sigma$ and $M_3 \le \tilde\kappa_{\max}(\tilde\alpha, n_\sigma, \mathfrak{C})^{3/4}$. The second condition holds because $M_3 \le \tilde\kappa^{3/4}$ by Jensen's Inequality (Lin and Bai, 2010, 8.4.b). Thus, in the two stage algorithm we have

$$\begin{aligned}
\Pr\left(|\hat\mu_n - \mu| \le \varepsilon\right) &= \mathbb{E}\bigl(\Pr\left(|\hat\mu_n - \mu| \le \varepsilon \mid \hat\sigma\right)\bigr) \\
&\ge \mathbb{E}\left((1-\tilde\alpha)1_{\sigma\le\hat\sigma}\right) \\
&\ge (1-\tilde\alpha)(1-\tilde\alpha) = 1-\alpha. \qquad \square
\end{aligned}$$

As mentioned in the introduction, one popular case is occurs when Y is a d-variate function of a random vector $\boldsymbol{X}$. In this case μ corresponds to the multivariate integral in (6) and Theorem 5 may be interpreted as below:

Corollary 1. *Suppose that $\rho : \mathbb{R}^d \to \mathbb{R}$ is a probability density, $f : \mathbb{R}^d \to \mathbb{R}$ has finite $\mathcal{L}_4$ norm as defined in* (7)*, and furthermore f lies in the cone $\mathcal{C}_{\tilde\kappa_{\max}}$ defined in* (8)*, where $\tilde\kappa_{\max} = \tilde\kappa_{\max}(\tilde\alpha, n_\sigma, \mathfrak{C})$. It follows that Algorithm 1 yields an estimate, $\hat\mu$, of the multidimensional integral μ defined in* (6)*, which satisfies the fixed width confidence interval*

$$\Pr(|\hat\mu - \mu| \le \varepsilon) \ge 1-\alpha.$$

3.4 Cost of the Algorithm

The number of function values required by the two-stage Algorithm 1 is $n_\sigma + n$, the sum of the initial sample size used to estimate the variance of Y and the sample size used to estimate the mean of Y. Since n is a random variable, the cost of this algorithm might best be defined defined probabilistically. Moreover, the cost depends strongly on σ^2, the unknown variance of Y, as well as the error tolerance (interval width) ε. We define the algorithmic cost for this problem to reflect these features.

Let A be any random algorithm that takes as its input, a method for generating random samples, $Y_1, Y_2, \ldots$ with common distribution function F having variance σ^2 and modified kurtosis $\tilde{\kappa}$, an error tolerance, ε, and an uncertainty tolerance, α. The algorithm then computes $\hat{\mu} = A(F, \varepsilon, \alpha)$, an approximation to $\mu = \mathbb{E}(Y)$, based on a total of $\text{cost}(A, \varepsilon, \alpha, F)$ samples. The probabilistic cost of the algorithm, with uncertainty β, for integrands of variance no greater than $\sigma^2_{\max}$ and modified kurtosis no greater than $\tilde{\kappa}_{\max}$ is defined as

$$\text{cost}(A, \varepsilon, \alpha, \beta, \tilde{\kappa}_{\max}, \sigma_{\max}) := \sup_{\substack{\tilde{\kappa} \le \tilde{\kappa}_{\max} \\ \sigma \le \sigma_{\max}}} \min \{N : \Pr[\text{cost}(A, \varepsilon, \alpha, F) \le N] \ge 1 - \beta\}. \tag{21}$$

The cost of the particular two-stage Monte Carlo algorithm defined in Algorithm 1, denoted TwoStage, is

$$\sup_{\substack{\tilde{\kappa} \le \tilde{\kappa}_{\max} \\ \sigma \le \sigma_{\max}}} \min \left\{ N : \Pr(n_\sigma + N_\mu(\varepsilon, \hat{\sigma}, \tilde{\alpha}, \tilde{\kappa}_{\max}^{3/4}) \le N) \ge 1 - \beta \right\}. \tag{22}$$

Since n_σ is fixed, bounding this cost depends on bounding $N_\mu(\varepsilon, \hat{\sigma}, \tilde{\alpha}, \tilde{\kappa}_{\max}^{3/4})$, which depends on $\hat{\sigma}$ as given by Algorithm 1. Moreover, $\hat{\sigma}$ can be bounded above using (13b) in Lemma 1. For $\tilde{\kappa} \le \tilde{\kappa}_{\max}$,

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

where

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

Noting that $N_\mu(\varepsilon, \cdot, \tilde{\alpha}, \tilde{\kappa}_{\max}^{3/4})$ is a non-decreasing function allows one to derive the following upper bound on the cost of the adaptive Monte Carlo algorithm.

Theorem 6. *The two stage Monte Carlo algorithm for fixed width confidence intervals based on IID sampling described in Algorithm 1, denoted* TwoStage*, has a probabilistic cost bounded above by*

$$\text{cost}(\text{TwoStage}, \varepsilon, \alpha, \beta, \mathscr{C}_{\kappa_{\max}}, \sigma_{\max}) \le n_\sigma + N_\mu(\varepsilon, \sigma_{\max} v(\tilde{\alpha}, \beta, \mathfrak{C})), \tilde{\alpha}, \tilde{\kappa}_{\max}^{3/4}).$$

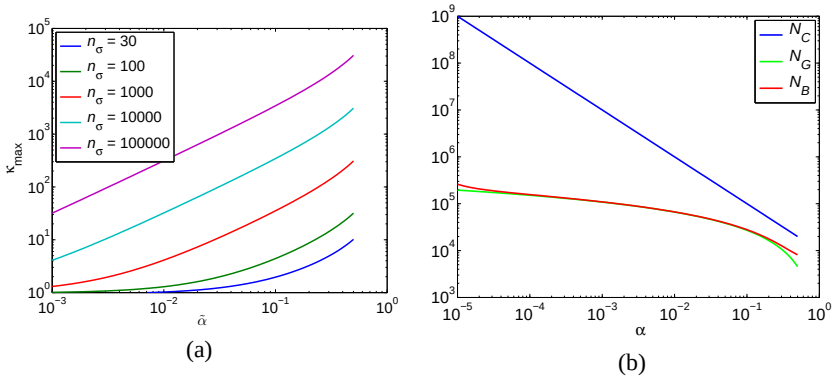
Note that the Chebychev sample size, N_{Cheb}, defined in (16), the Berry-Esseen sample size, N_{BE}, defined in (18), and thus N_μ all depend on σ and ε through their ratio, σ/ε. Apart from

The cost of the adaptive Monte Carlo algorithm TwoStage is roughly proportional to $\sigma_{\max}^2\varepsilon^{-2}$. The set $\mathscr{C}_{\kappa_{\max}}$ contains integrands with arbitrarily large variance, $\sigma^2 = \text{Var}(f)$, and thus with potentially arbitrarily large algorithmic cost. On the other hand, since the algorithm is adaptive, the cost may be small if σ^2 is small. The upper bound in Theorem 6 certainly scales with the $\sigma_{\max}^2$ as one might hope if $\sigma_{\max}^2$ were known. The variable cost of the algorithm for integrands in $\mathscr{C}_{\kappa_{\max}}$ 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 its kurtosis, κ, which is unchanged when the integrand is multiplied by an arbitrary nonzero constant.

Though n_σ can be as small as 2 it should ordinarily be larger.

As shown in Figure 1b, for a range of α, the sample size guaranteeing coverage of the confidence interval, N_B, is quite close to the sample size for the approximate Central Limit Theorem confidence interval, N_G, however, N_B may be somewhat larger for very small or rather large α. In general N_B is significantly smaller than N_C, but not always. A disadvantage of (**??**) is that class of integrands, $\mathscr{L}_3$, is smaller than that in (**??**), but this typically a small price to pay given the much smaller cost of computation.

Figure 1a shows how large a kurtosis can be accommodated for a given n_σ, α, and $\mathfrak{C} = 1.5$. Note that for $n = 30$, a common rule of thumb for applying the central limit theorem, even the modest value $\alpha = 0.1$ yields $\tilde{\kappa}_{\max}$ of only about 2, corresponding to a kurtosis of about 5. While that kurtosis is reasonably large for many observational data settings, Monte Carlo applications often involve a larger kurtosis than this.



(a) (b)

Fig. 1. (a) The maximum kurtosis, $\kappa_{\max}(\alpha, n_\sigma, 1.5)$, as defined in (15); (b) comparison of sample sizes $N_G(0.01,\alpha)$, $N_C(0.01,\alpha)$, and $N_B(0.01,\alpha,\kappa_{\max}^{3/4}(\alpha,1000,1.5))$.

Remark 1. If one is willing to invest n_σ samples to estimate σ, it makes practical sense to to choose the sample size for the sample mean at least that large, i.e,

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

By the error bound following from Chebychev's inequality, (**??**), this means that the probabilistic absolute error bound in Theorem 5 also holds for integrands, f, lying in the ball $\mathscr{B}_{\sigma_{\max}}$, defined in (**??**), where $\sigma_{\max} = \varepsilon\sqrt{\alpha n_\sigma}$.

4 Numerical Examples with TwoStage

4.1 Illustrative Univariate Examples of Automatic Algorithms

Several commonly used software packages have automatic algorithms for integrating functions of a single variable. These include

- `quad` in MATLAB (The MathWorks, Inc., 2012), adaptive Simpson's rule based on `adaptsim` by Gander and Gautschi (2000),
- `quadgk` in MATLAB (The MathWorks, Inc., 2012), adaptive Gauss-Kronrod quadrature based on `quadva` by Shampine (2008), and
- the `chebfun` (Hale et al, 2012) toolbox for MATLAB (The MathWorks, Inc., 2012), which approximates integrals by integrating interpolatory Chebychev polynomial series for the integrands.

For the first three of these automatic algorithms one can easily probe where they sample the integrand, feed the algorithms zero values, and then construct fooling functions that the automatic algorithms will return a zero value for the integral. Figure 2 displays these fooling functions for the problem $\mu = \int_0^1 f(x)\,\mathrm{d}x$ for the first three algorithms. Each of these algorithms is asked to provide an answer with an absolute error no greater than 10^{14}, but in fact the the absolute error is 1 for these fooling functions. The algorithms `quad` and `chebfun` sample only about a dozen points before concluding that the function is zero, whereas the algorithm `quadgk` samples a much larger number of points (only those between 0 and 0.01 are shown in the plot). Algorithm `NIntegrate` is hard snoop, but it is examined for the next example.

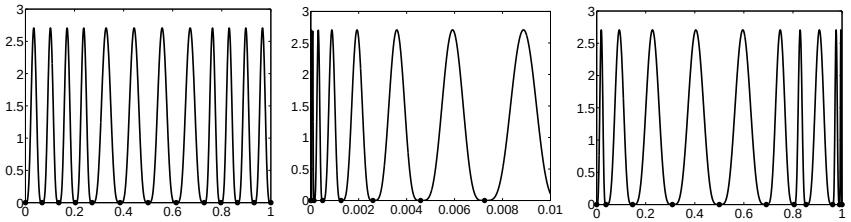


Fig. 2. Plots of fooling functions, f, with $\mu = \int_0^1 f(x)\,\mathrm{d}x = 1$, but for which the corresponding algorithms return values of $\hat{\mu} = 0$.

Accuracy and timing results have been recorded for the test function

$$f(\boldsymbol{x}) = a_0 + b_0 \prod_{j=1}^{d} \left[1 + b_j \exp\left(-\frac{(x_j - h_j)^2}{c_j^2} \right) \right]. \tag{23}$$

Here $\boldsymbol{x}$ is a d dimensional vector, and the b_j, c_j, and h_j are parameters. For Figure 3 shows the results of different algorithms being used to integrate 500 different instances of f. For each instance of f, the parameters are chosen as follows:

- $b_j \in [0.1, 10]$ with $\log(b_j)$ being i.i.d. uniform, $j = 1, \ldots, d$,
- $c_j \in [10^{-6}, 1]$ with $\log(c_j)$ being i.i.d. uniform, $j = 1, \ldots, d$,
- $h_j \in [0, 1]$ with h_j being i.i.d. uniform, $j = 1, \ldots, d$,
- b_0 chosen in terms of the b_j, c_j, and h_j in order to make $\sigma^2(f) \in [10^{-2}, 10^2]$, with $\log(\sigma(f))$ being i.i.d. uniform for each instance, and
- a_0 in terms of chosen in terms of the b_j, c_j, and h_j to make $\mu(f) = 1$.

These 500 random constructions of f with $d = 1$ are integrated over $[0, 1]$ with $\rho = 1$ (the uniform density function), using `quad`, `quadgk`, and `chebfun`. For the first two of these algorithms, the specified absolute error tolerance is $\varepsilon = 0.001$. The algorithm `chebfun` attempts to do all calculations to near machine precision. The observed error and execution times are plotted in Figure 3. Whereas `chebfun` uses a minimum of $2^3 + 1 = 9$ function values, the figure labeled "`chebfun` heavy duty" displays the results of requiring chebfun to use at least $2^{10} + 1 = 1025$ function values.

Figure 3 shows that `quad` and `quadgk` are quite fast, nearly always providing an answer in less that 0.01 seconds. Unfortunately, they successfully meet the error tolerance only about 30% of the time for `quad` and 50–60% of the time for `quadgk`. The difficult cases are those where c_1 is quite small, and these algorithms miss the sharp peak. The performance of `chebfun` is similar to that of `quad` and `quadgk`. The heavy duty version of `chebfun` fares somewhat better.

Figure 3 shows timing and observed errors for the adaptive algorithm, TwoStage, with i.i.d. sampling, as described in the previous section. The parameters chosen are $\varepsilon = 0.001$, $\alpha = 5\%$, and $\mathfrak{C} = 1.5$. For the plot on the left, $n_\sigma = 2^{10} = 1024$, which corresponds to $\kappa_{\max} = 9.2$. For the heavy duty plot on the right, $n_\sigma = 2^{17} = 131\,072$, which corresponds to $\kappa_{\max} = 1052$. In both of these plots the points labeled with a * are those for which $\kappa(f) \le \kappa_{\max}$ and so Theorem 5 guarantees that the answer is correct $1 - \alpha = 95\%$ of the time. For these plots all of the points labeled * fall within the prescribed error tolerance. For TwoStage i.i.d. heavy duty plot $\kappa_{\max}$ is larger, so there are more points for which the guarantee holds. Those points labeled with a dot, are those for which $\kappa(f) > \kappa_{\max}$, and so no guarantee holds. The points labeled with a diamond are those for which TwoStageattempts to exceed the cost budget, i.e., it wants to choose n such that $(n_\sigma + n)d > N_{\max} = 10^9$.

The `cubMC` algorithm performs somewhat more robustly than `quad`, `quadgk`, and `chebfun`, because TwoStagedoes requires a very low degree of smoothness and makes a fairly large minimum sample. The more important point is that TwoStagehas a guarantee, where to our knowledge, the other routines do not.

Figure 3 also exhibits the results of using TwoStagewith scrambled Sobol' sampling (Owen, 1995, 1997a,b; Matoušek, 1998; Hong and Hickernell, 2003; Dick and Pillichshammer, 2010), i.e., the sample mean, $\hat{\mu}_n$ is based on sampling the integrand on a Sobol' net with n points. Since the points of the Sobol' net are purposefully correlated, the error of $\hat{\mu}_n$ does not depend on $\mathrm{Var}(f)$, but on some measure of variation of f (see (Owen, 1995, 1997a,b; Dick and Pillichshammer, 2010)) that is difficult to estimate in practice. Tony Warnock and John Halton proposed the following error bound estimate, called the quasi-standard error (Halton, 2005; Owen, 2006):

$$\mathrm{qse}(f; m) = \sqrt{\frac{1}{m(m-1)} \sum_{j=1}^{m} [\hat{\mu}_n^{(j)} - \hat{\mu}_n]^2}. \tag{24}$$

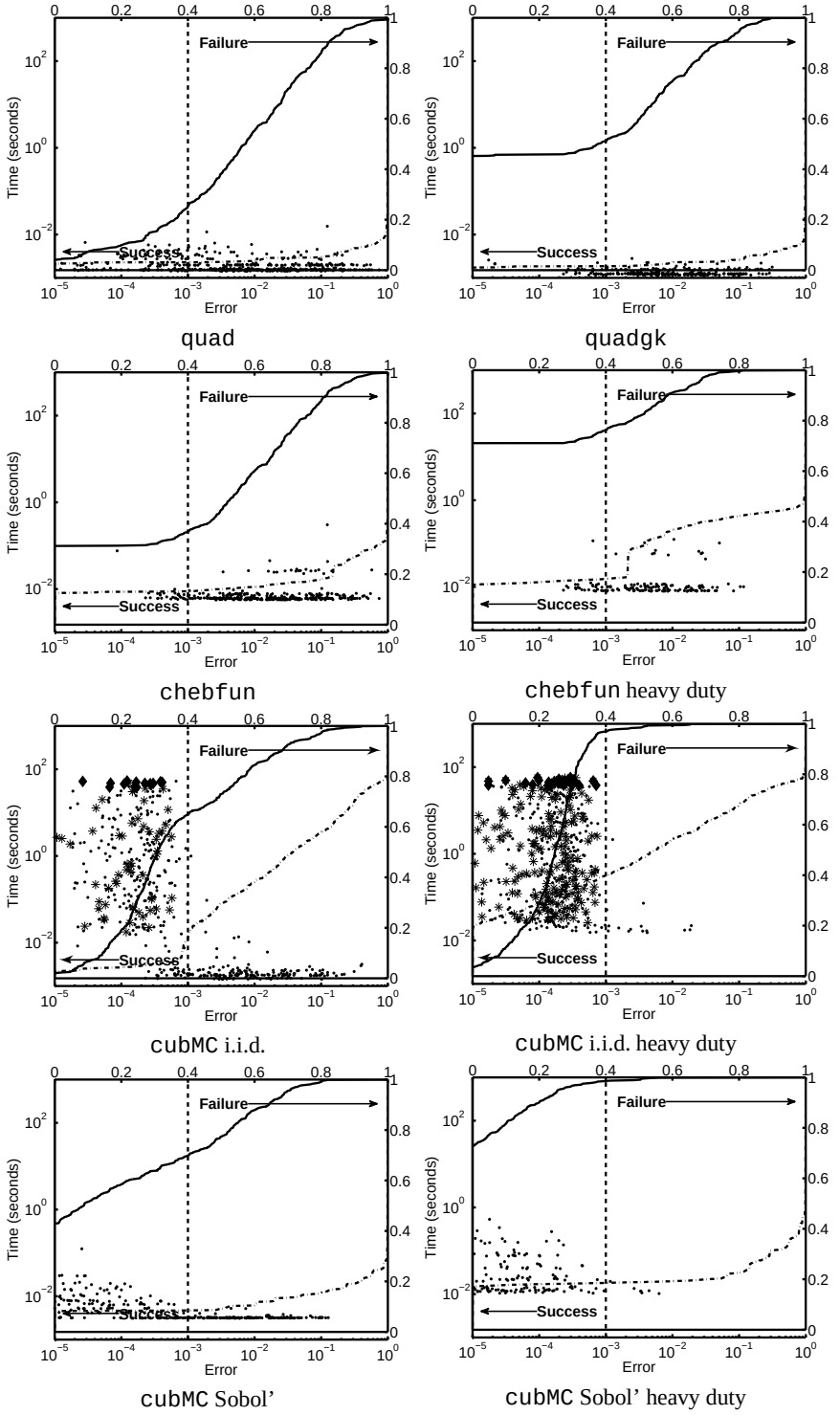


Fig. 3. Execution times and errors for test function (23) for $d = 1$ and $\varepsilon = 10^{-3}$, and a variety of parameters giving a range of $\sigma(f)$ and $\kappa(f)$. The solid line shows that cumulative distribution of actual errors, and the dot-dashed line shows the cumulative distribution of execution times. For the `cubMC` i.i.d. and i.i.d. heavy duty the points labeled * are those for which the Theorem 5 guarantees the error tolerance.

where $\hat{\mu}^{(j)}$ corresponds to the sample mean of the function values for the j^{th} partition out of the Sobol' net. In TwoStagethe error of $\hat{\mu}_n$ is assumed to be no greater than $\mathfrak{C}\text{qse}(f;m)$ with $m=8$ and $\mathfrak{C}=1.5$. The number of samples, n, is doubled until $\mathfrak{C}\text{qse}(f;m)\le\varepsilon$. Unfortunately, there is no theory yet that intuitively describes the cone of integrands for which this stopping criterion guarantees that the error tolerance is met. This is an area of ongoing research.

Clearly, from Figure 3, the Sobol' sampling option is more reliable and takes less time than the i.i.d. option of TwoStage. This is due primarily to the fact that in dimension one, Sobol' sampling is equivalent to stratified sampling, where the points are more evenly spread.

Figure 4 repeats the simulation shown in Figure 3 for the same test function (23), but now with $d=2,\ldots,8$ chosen randomly and uniformly. For this case the univariate integration algorithms are inapplicable, but TwoStagewith both sampling schemes, i.i.d. and Sobol', can be used. There are more cases where the TwoStagetries to exceed the maximum sample size allowed, but the behavior seen for $d=1$ still generally apply.

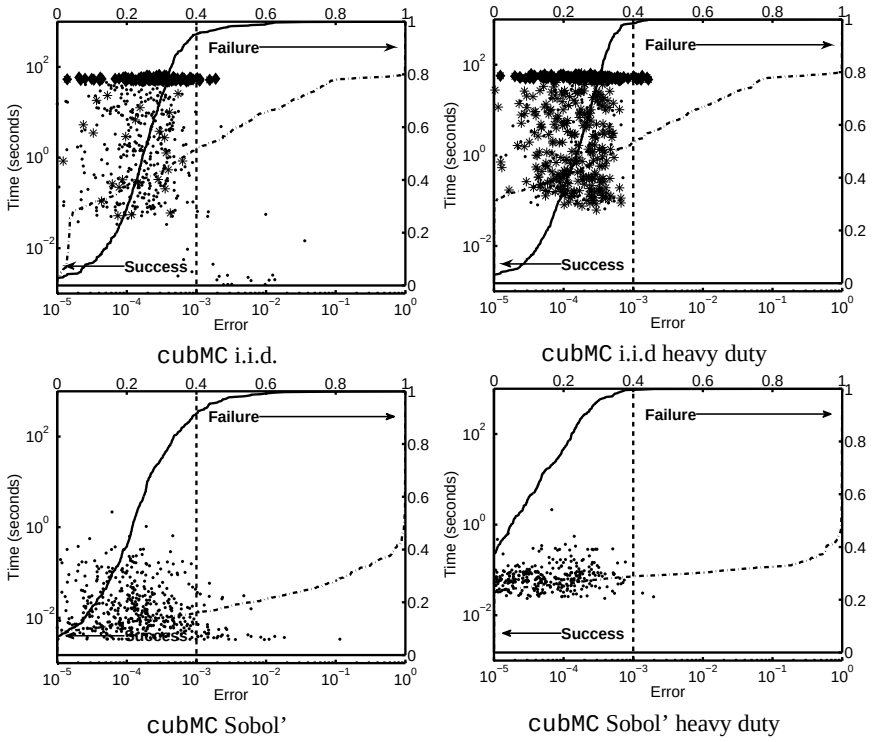


Fig. 4. Execution times and errors for test function (23) for $d=2,\ldots,8$ and $\varepsilon=10^{-3}$, with the rest of the parameters as in Figure 3.

We will put one more finance example here.

5 A General 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

$$\Pr\left[\left|\frac{\tilde{\mu}-\mu}{\mu}\right| \leq \varepsilon\right] \geq 1-\alpha.$$

A more general form of this criterion would be

$$\Pr\left[\frac{|\tilde{\mu}-\mu|}{1-\theta+\theta|\mu|} \leq \varepsilon\right] \geq 1-\alpha. \tag{25}$$

for some fixed $\theta \in [0,1]$, where $\theta = 0$ corresponds to absolute error, and $\theta = 1$ corresponds to relative error. Clearly, one must have $(1-\theta)+|\mu| \neq 0$ for such a statement to be possible.

If $\varepsilon_A \geq 0$ is an absolute error tolerance, and $\varepsilon_R \geq 0$ is a relative error tolerance, then letting

$$\varepsilon = \frac{\varepsilon_A \varepsilon_R}{\theta\varepsilon_A + (1-\theta)\varepsilon_R},$$

it follows that for all $\theta \in [0,1]$,

$$[1-\theta+\theta|\mu|]\varepsilon = (1-\gamma)\varepsilon_A + \gamma\varepsilon_R|\mu| \leq \max(\varepsilon_A, \varepsilon_R|\mu|),$$

where

$$\gamma = \frac{\theta\varepsilon_A}{\theta\varepsilon_A + (1-\theta)\varepsilon_R} \in [0,1].$$

Thus, error criterion (25) implies that one has satisfied either an absolute or a relative error criterion,

$$\Pr\left[\frac{|\tilde{\mu}-\mu|}{1-\theta+\theta|\mu|} \leq \varepsilon\right] \geq 1-\alpha \implies \Pr[|\tilde{\mu}-\mu| \leq \varepsilon_A \text{ or } |\tilde{\mu}-\mu| \leq |\mu|\varepsilon_R] \geq 1-\alpha.$$

A value of γ close to zero implies a preference to fulfill the absolute error criterion while a value of γ close to one implies a preference to fulfill the relative error criterion.

Obtaining a confidence interval of the form (25), proceeds in three stages: i) obtaining an upper bound on σ^2, ii) obtaining a lower bound on $1-\theta+\theta|\mu|$, and iii) then using these to obtain (25). What differs from the absolute error case is step ii). For this step it is noted that

$$\begin{aligned}\Pr[|\hat{\mu}-\mu| \leq \hat{\varepsilon}] \geq 1-\alpha &\implies \Pr[\max(|\hat{\mu}|-\hat{\varepsilon},0) \leq |\mu| \leq |\hat{\mu}|+\hat{\varepsilon}] \geq 1-\alpha \\ &\implies \Pr[1-\theta+\theta\max(|\hat{\mu}|-\hat{\varepsilon},0) \leq 1-\theta+\theta|\mu| \leq 1-\theta+\theta(|\hat{\mu}|+\hat{\varepsilon})] \geq 1-\alpha. \end{aligned} \tag{26}$$

Although one might be happy the left side of this inequality being positive, if it is too much smaller than the right side, then one might be eventually expending too much extra work in step iii). Thus, it makes sense to require

$$\begin{aligned} &1-\theta+\theta\max(|\hat{\mu}|-\hat{\varepsilon},0) \geq \hat{\delta}[1-\theta+\theta(|\hat{\mu}|+\hat{\varepsilon})] \\ \iff \hat{\varepsilon} = &\begin{cases} \dfrac{(1-\hat{\delta})(1-\theta)}{\hat{\delta}\theta} - |\hat{\mu}|, & 0 \leq |\hat{\mu}| < \dfrac{(1-\hat{\delta})(1-\theta)}{2\hat{\delta}\theta}, \\ \dfrac{1-\hat{\delta}}{1+\hat{\delta}}\left[\dfrac{1-\theta}{\theta} + |\hat{\mu}|\right], & \dfrac{(1-\hat{\delta})(1-\theta)}{2\hat{\delta}\theta} \leq |\hat{\mu}| < \infty. \end{cases} \end{aligned}$$

This is done iteratively in the algorithm described in Theorem 7 below. One needs to prevent $\hat{\varepsilon}$ from becoming too small. This means that $\hat{\delta}$ should kept away from 1, which means that the lower bound on $1-\theta+\theta|\mu|$ is allowed to be somewhat smaller than the upper bound. Preventing $\hat{\varepsilon}$ from becoming too small also means that $1-\theta+|\hat{\mu}|$ cannot be too small. This may be unavoidable if one is interested in relative error $\theta=1$, and the true answer, μ, is small.

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 $N_{CB}(\cdot,\alpha,M)$,

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

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

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

It then follows then by Chebychev's inequality and the Berry-Esseen Inequality (see (17)) that

$$\Pr[|\hat{\mu}_n-\mu|<\hat{\varepsilon}] \ge 1-\alpha, \quad \text{provided } \ f \in \mathscr{C}_{\kappa_{\max}}, \text{ where } \hat{\varepsilon} = \sigma(f) N_{CB}^{-1}(n,\alpha,\kappa_{\max}^{3/4}),$$

and $\sigma(f) = \sqrt{\mathrm{Var}(f)}$ is the standard deviation of the integrand. Given a significance level, $\alpha \in (0,1)$, let $\alpha_\sigma, \alpha_\mu, \alpha_1, \alpha_2, \ldots$ be an infinite sequence of positive numbers all less than one, such that

$$(1-\alpha_\sigma)(1-\alpha_\mu)(1-\alpha_1)(1-\alpha_2)\cdots = 1-\alpha. \tag{27}$$

For example, one might choose $\alpha_\sigma, \alpha_\mu$, and $\hat{\alpha}$ such that $(1-\alpha_\sigma)(1-\alpha_\mu)(1-\hat{\alpha}) = 1-\alpha$, and then

$$\alpha_i = 1-(1-\hat{\alpha})^{(a-1)a^{-i}}, \ i \in \mathbb{N}, \quad \text{where } a \in (1,\infty). \tag{28}$$

Theorem 7. *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,* $\mathfrak{C} \in (1,\infty)$,
- *factors for Step 2,* $\hat{\delta}, \delta, \tilde{\delta} \in (0,1)$, *with* $\delta < \tilde{\delta}$.
- *uncertainty tolerance,* $\alpha \in (0,1)$, *and a sequence* $\alpha_\sigma, \alpha_\mu, \alpha_1, \alpha_2, \ldots$ *satisfying* (27),
- *the parameter* $\theta \in [0,1]$, *used to define the general error criterion* (25), *and*
- *the error tolerance,* $\varepsilon > 0$.

Let $\kappa_{\max} = \kappa_{\max}(n_\sigma, \alpha_\sigma, \mathfrak{C})$ *as defined in* (15). *For any* f *lying in the cone of functions with bounded kurtosis,* $\mathscr{C}_{\kappa_{\max}}$, *do the following:*

1. **Bounding the variance of the integrand from above.** *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* (**??**). *Compute the width of initial the confidence interval for the mean,* $\hat{\varepsilon}_1 = \hat{\sigma} N_{CB}^{-1}(n_1, \alpha_1, \kappa_{\max}^{3/4})$.
2. **Bounding the denominator in the error criterion from below.** *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) *Compute* $\mathfrak{C} = 1-\theta+\theta\max(|\hat{\mu}_{n_i}|-\hat{\varepsilon}_i, 0)$, *a confident lower bound on* $1-\theta+\theta|\mu|$, *according to* (26). *If* $\mathfrak{C} \ge \delta[1-\theta+\theta(|\hat{\mu}_{n_i}|+\hat{\varepsilon})]$, *then* $\mathfrak{C}$ *is large enough. Set* $\tau = i$ *and go to Step 3.*

c) Else, compute the next tolerance for the sample mean

$$\hat{\varepsilon}_0 = \begin{cases} \dfrac{(1-\hat{\delta})(1-\theta)}{\hat{\delta}\theta} - |\hat{\mu}_{n_i}|, & 0 \le |\hat{\mu}_{n_i}| < \dfrac{(1-\hat{\delta})(1-\theta)}{2\hat{\delta}\theta}, \\ \dfrac{1-\hat{\delta}}{1+\hat{\delta}}\left[\dfrac{1-\theta}{\theta} + |\hat{\mu}_{n_i}|\right], & \dfrac{(1-\hat{\delta})(1-\theta)}{2\hat{\delta}\theta} \le |\hat{\mu}_{n_i}| < \infty, \end{cases}$$

$$\hat{\varepsilon}_{i+1} = \max(\min(\hat{\varepsilon}_0, \tilde{\delta}\hat{\varepsilon}_i), \delta\hat{\varepsilon}_i).$$

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

3. **Computing the sample mean to sufficient accuracy.** *Compute the sample size $n = N_{CB}(\mathfrak{C}\varepsilon/\hat{\sigma}, \alpha_\mu, \kappa_{\max}^{3/4})$. Compute $\tilde{\mu} = \hat{\mu}_n$ using a simple random sample that is independent of those used to compute $\hat{v}_{n_\sigma}$ and $\hat{\mu}_{n_1}, \ldots, \hat{\mu}_{n_\tau}$. Terminate the algorithm.*

If this algorithm terminates, then the general error criterion, (25)*, 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 $n_1, \ldots, n_\tau, n$, the number of iterations, τ, required to get a good lower bound $\mathfrak{C}$, and the final estimate of the mean $\tilde{\mu} = \hat{\mu}_n$. 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,

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

Remark 2. Step 2 in this algorithm is not needed for the case of pure absolute error $\theta = 0$, because $\mathfrak{C} = 1$ automatically, which is large enough. As suggested earlier, a difficulty may arise if $\mu \approx 0$ and $\theta \approx 1$, in which the algorithm may fail to converge in a reasonable number of steps and overall sample size. Step 2c has safeguards against making $\hat{\varepsilon}_{i+1}$ too small compared to $\hat{\varepsilon}_i$, but this may also increase the number of iterations, τ, necessary for completion of Step 2. Because it is difficult to knowing how large τ is for a given integrand, there is no rigorous bound on the cost of this algorithm yet.

6 Discussion

Put something here.

Looking for algorithms that work well for cones of integrands, $\mathcal{C}_{\kappa_{\max}}$, leads one to *adaptive* algorithms. The sample size used to estimate the integral is determined adaptively by first computing an upper bound on $\|f - \mu(f)\|_2$. In information-based complexity theory it is known that adaptive information does not help for convex sets of integrands in the worst case

and probabilistic settings (Traub et al, 1988, Chapter 4, Theorem 5.2.1; Chapter 8, Corollary 5.3.1). Here, the cone, $\mathcal{C}_{\kappa_{\max}}$ is not a convex set, so adaption can help.

Again, it should be stressed that the algorithm to be presented here is automatic. It does not require information about $\|f-\mu(f)\|_2=\sigma$, but this quantity needs to be reliably estimated by the algorithm. Thus, the sample size needed, and consequently the time required, to estimate μ to within the prescribed error tolerance depends on how large $\|f-\mu(f)\|_2=\sigma$ is estimated to be. The algorithm is adaptive, and its cost depends on the integrand.

Acknowledgements

The authors gratefully acknowledge discussions with Erich Novak and Henryk Woźniakowski. The plots of the univariate fooling functions were prepared with the help of Nicholas Clancy and Caleb Hamilton.

References

Bahadur RR, Savage LJ (1956) The nonexistence of certain statistical procedures in nonparametric problems. Ann Math Stat 27:1115–1122

Dick J, Pillichshammer F (2010) Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration. Cambridge University Press, Cambridge

Gander W, Gautschi W (2000) Adaptive quadrature — revisited. BIT 40:84–101

Hale N, Trefethen LN, Driscoll TA (2012) Chebfun Version 4

Hall P (1986) On the bootstrap and confidence intervals. The Annals of Statistics 14:1431–1452

Halton JH (2005) Quasi-probability: Why quasi-Monte-Carlo methods are statistically valid and how their errors can be estimated statistically. Monte Carlo Methods and Appl 11:203–350

Hong HS, Hickernell FJ (2003) Implementing scrambled digital nets. ACM Trans Math Software 29:95–109

Lemieux C (2009) Monte Carlo and quasi-Monte Carlo Sampling. Springer Science+Business Media, Inc., New York

Lin Z, Bai Z (2010) Probability Inequalities. Science Press and Springer-Verlag, Beijing and Berlin

Matoušek J (1998) On the L_2-discrepancy for anchored boxes. J Complexity 14:527–556

McCullagh P (1987) Tensor methods in statistics. Chapman and Hall, London

Miller R (1986) Beyond ANOVA, Basics of Applied Statistics. John Wiley & Sons, Inc., New York

Niederreiter H (1992) Random Number Generation and Quasi-Monte Carlo Methods. CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia

Owen AB (1995) Randomly permuted (t,m,s)-nets and (t,s)-sequences. In: Niederreiter H, Shiue PJS (eds) Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, Springer-Verlag, New York, Lecture Notes in Statistics, vol 106, pp 299–317

Owen AB (1997a) Monte Carlo variance of scrambled net quadrature. SIAM J Numer Anal 34:1884–1910

Owen AB (1997b) Scrambled net variance for integrals of smooth functions. Ann Stat 25:1541–1562

Owen AB (2006) On the Warnock-Halton quasi-standard error. Monte Carlo Methods and Appl 12:47–54

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

Shampine LF (2008) Vectorized adaptive quadrature in matlab. J Comput Appl Math 211:131–140

Siegmund D (1985) Sequential Analysis: Tests and Confidence Intervals. Springer, New York

Sloan IH, Joe S (1994) Lattice Methods for Multiple Integration. Oxford University Press, Oxford

The MathWorks, Inc (2012) MATLAB 7.12. The MathWorks, Inc., Natick, MA

Traub JF, Wasilkowski GW, Woźniakowski H (1988) Information-Based Complexity. Academic Press, Boston