ADAPTIVE SIMPLE MONTE CARLO

FRED J. HICKERNELL, LAN JIANG, AND YUEWEI LIU

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

1. 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. 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 Monte Carlo in Practice

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:

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

A natural question is the following: *How large should n be to ensure with some reasonable certainty, that the error of the approximation is no larger than some error tolerance, ε?* In practice, a common approach is to invoke the Central Limit Theorem, which says that for a significance level or uncertainty tolerance, α, one has

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

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

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

This variance σ^2 can be estimated in terms of the sample variance based on an initial sample of size, $n_0 \in \mathbb{N}$:

$$\hat{\sigma}^2 = \mathfrak{C}^2 \hat{v}_{n_0}, \qquad \text{where} \quad \hat{v}_{n_0} = \frac{1}{n_0 - 1} \sum_{i=1}^{n_0} [Y_i - \hat{\mu}_n]^2, \quad Y_i = f(\boldsymbol{X}_i). \tag{3}$$

The fudge 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

$$N_G(\varepsilon, \alpha) := \left\lceil \left(\frac{z_{\alpha/2}}{\varepsilon}\right)^2 \right\rceil, \tag{4}$$

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:

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=n_0+1}^{n_0+n} Y_i, \qquad Y_i = f(\boldsymbol{X}_i). \tag{5}$$

This algorithm has a cost of $n_0 + N_G(\varepsilon/\hat{\sigma}, \alpha)$ function evaluations, which depends on the initial sample size, n_0, the error tolerance, ε, the degree of uncertainty, α, the fudge factor, $\mathfrak{C}$, and the sample variance, $\hat{\sigma}^2$.

The algorithm defined by (12), (4), and (5) 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.

3. Adaptive Monte Carlo

First the question of bounding the variance is addressed. The integrands under consideration are assumed to have finite moments of up to a certain order. Define the $\mathcal{L}_p$ norm of f as follows:

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

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

$$M_p(f) := \int_{\mathbb{R}^d} |f(\boldsymbol{x}) - \mu|^p \, \rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \|f - \mu\|_p^p, \qquad p \ge 1 \tag{6}$$

By these definitions, the variance of Y_1 is the second moment, i.e., $M_2(f) = \text{var}(Y_1) = \sigma^2$, and all functions in $\mathcal{L}_2$ have finite variance.

As mentioned in (12), a practical upper bound on this variance is obtained from the sample variance, $\hat{\sigma}^2 = \mathfrak{C}^2 \hat{v}_{n_0}$, where $\mathfrak{C} > 1$ is some fudge factor. This can be justified in a probabilistic sense by appealing to Cantelli's inequality (Theorem 6) and the variance of $\hat{v}_{n)}$ given by Theorem 5. Proposition 7 implies that

$$\text{Prob}\left[\frac{\hat{v}_{n_0}}{1 - \sqrt{\left(\kappa - \frac{n_0-3}{n_0-1}\right)\left(\frac{1-\alpha}{\alpha n_0}\right)}} > \sigma^2\right] \ge 1 - \alpha,$$

where κ is the *kurtosis* of the integrand:

$$\kappa := \text{kurt}(f) = \frac{M_4(f)}{\text{var}^2(f)} = \frac{M_4(f)}{\sigma^4} \ge 1 \qquad \forall f \in \mathcal{L}_4. \tag{7}$$

Note that the kurtosis of a function is independent of its scale. Moreover, functions in $\mathcal{L}_4$ automatically have finite variance. It follows that the kurtosis of the integrand must be small enough, relative to n_0 and $\mathfrak{C}$ to ensure that the variance estimate is correct in a probabilistic sense, namely,

$$\frac{1}{1 - \sqrt{\left(\kappa - \frac{n_0-3}{n_0-1}\right)\left(\frac{1-\alpha}{\alpha n_0}\right)}} \le \mathfrak{C}^2$$

$$\frac{1}{\mathfrak{C}^2} \le 1 - \sqrt{\left(\kappa - \frac{n_0-3}{n_0-1}\right)\left(\frac{1-\alpha}{\alpha n_0}\right)}$$

$$\left(\kappa - \frac{n_0-3}{n_0-1}\right)\left(\frac{1-\alpha}{\alpha n_0}\right) \le \left(1 - \frac{1}{\mathfrak{C}^2}\right)^2$$

$$\kappa \le \frac{n_0-3}{n_0-1} + \left(\frac{\alpha n_0}{1-\alpha}\right)\left(1 - \frac{1}{\mathfrak{C}^2}\right)^2 =: \kappa_{\max}(\alpha, n_0, \mathfrak{C}). \tag{8}$$

Figure 4 shows how large a kurtosis can be accommodated for a given n, α, and fudge 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.

The other issue that needs to be addressed is a tight probabilistic error bound. The error bound given by the Central Limit Theorem, (1), 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

$$\text{Prob}\left[|\hat{\mu}_n - \mu| \le \varepsilon\right] \ge 1 - \alpha \quad \text{for } n \ge N_C(\varepsilon/\sigma, \alpha), \qquad \text{where } N_C(\varepsilon, \alpha) := \left\lceil \frac{1}{\alpha \varepsilon^2} \right\rceil. \tag{9}$$

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

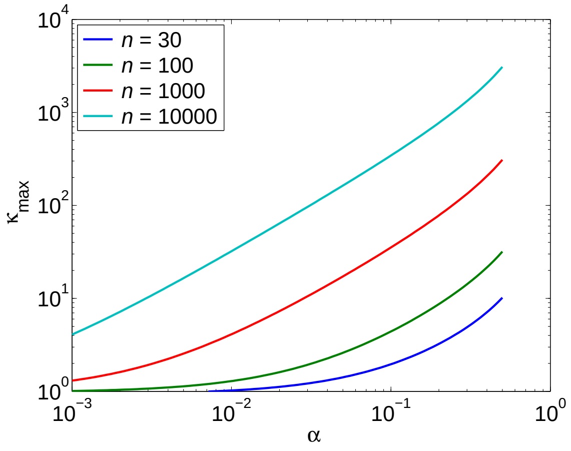

FIGURE 1. The maximum kurtosis, $\kappa_{\max}(n, \alpha, 1.5)$, as defined in (14).

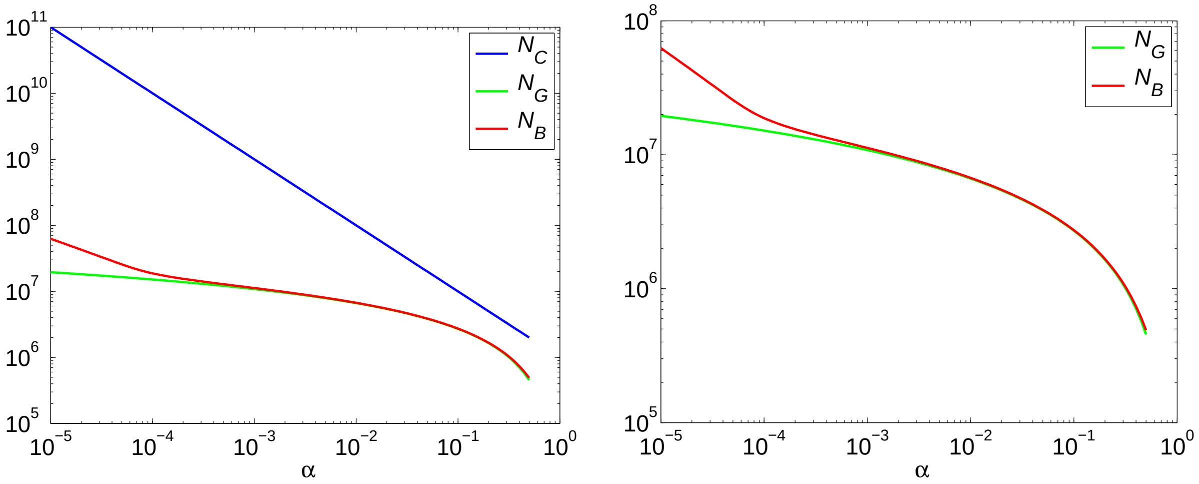

FIGURE 2. Comparison of $N_G(\varepsilon, \alpha)$, $N_C(\varepsilon, \alpha)$, and $N_B(\varepsilon, \alpha, \varrho)$ for $\varepsilon = 0.001$, and $\varrho = 5$.

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 (9) with a rigorous probabilistic bound can be found by invoking the non-uniform Berry-Esseen inequality (Theorem 4). This inequality makes strong assumptions on the distribution of $f(\boldsymbol{X})$, namely, a finite third moment, $M_3 < \infty$ 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

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

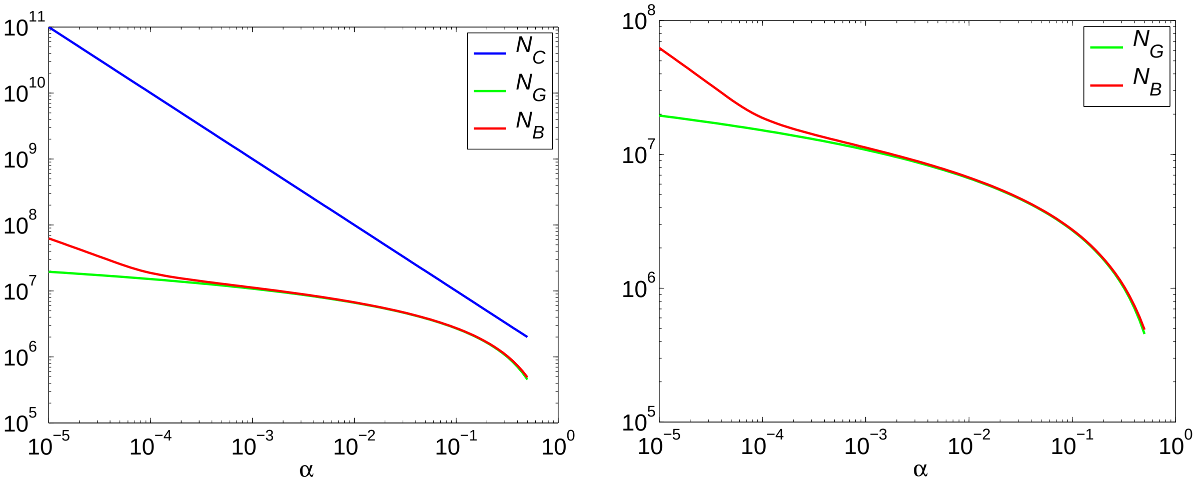

FIGURE 3. Comparison of $N_G(\varepsilon,\alpha)$, $N_C(\varepsilon,\alpha)$, and $N_B(\varepsilon,\alpha,\varrho)$ for $\varepsilon = 0.001$, and $\varrho = 5$.

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

$$\text{Prob}[|\hat{\mu}_n - \mu| < \varepsilon] \geq 1-\alpha, \quad \text{provided } n \geq N_B(\varepsilon/\sigma, \alpha, \varrho), \tag{10a}$$

where

$$N_B(b,\alpha,\varrho) := \min\left\{ n \in \mathbb{N} : \Phi\left(-b\sqrt{n}\right) + \frac{A\varrho}{\sqrt{n}}\left(1+b\sqrt{n}\right)^{-3} \leq \frac{\alpha}{2} \right\}. \tag{10b}$$

Definition (10b) may be re-written implicitly

$$N_B = \left\lceil \left(\frac{z_{\alpha/2 - A\varrho/(\sqrt{N_B}(1+b\sqrt{N_B})^3)}}{b} \right)^2 \right\rceil. \tag{11}$$

The definition of N_B implies that

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

Here A is some absolute number .... As shown in Figure 3, N_B is close to N_G for moderate α, but N_B may be significantly larger for very small α. In general N_B is smaller than N_C. The disadvantage of (**??**) is that class of integrands is smaller than that in (**??**), but this typically a small price to pay given the much smaller cost of computation.

The error analyses of the simple, non-adaptive Monte Carlo method has some practical shortcomings. To apply the randomized or probabilistic cases one must have some a priori knowledge of σ^2, the variance of the integrand. It must not larger than $\sigma^2_{\max}$ or the theory does not apply. If the variance of the integrand is much smaller than $\sigma^2_{\max}$, then the specified sample sizes,

$$\left\lceil \frac{\sigma^2_{\max}}{\varepsilon^2} \right\rceil \text{ in (??)}, \qquad N_C(\varepsilon/\sigma_{\max},\alpha) \text{ in (??), or} \qquad N_B(\varepsilon/\sigma_{\max},\alpha,\varrho_{\max}) \text{ in (??)},$$

are much too large. To overcome this shortcoming requires an adaptive Monte Carlo algorithm that estimates the sample size based on the sample variance, as suggested in the introduction.

4. Adaptive Monte Carlo

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

$$\hat{v}_n = \frac{1}{n-1}\sum_{i=1}^{n}[f(\boldsymbol{X}_i) - \hat{\mu}_n]^2. \tag{12}$$

To ensure that one does not underestimate the variance, and thus the needed sample size, one might choose to approximate the variance by $\hat{\sigma}_n^2 = L^2\hat{v}_n$, where $L > 1$ is some fudge factor. This can be formalized in a probabilistic sense by appealing to Cantelli's inequality (Theorem 6) and the variance of $\hat{v}_n$ in Theorem 5. Proposition 7 implies that

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

where

$$\kappa := \text{kurt}(f) = \frac{M_4(f)}{\text{var}^2(f)} = \frac{M_4(f)}{\sigma^4} \ge 1 \qquad \forall f \in \mathcal{L}_4 \tag{13}$$

denotes the *kurtosis*. Note that the kurtosis of a function is independent of scale. Moreover, functions in $\mathcal{L}_4$ automatically have finite variance. It follows that the kurtosis of the integrand must be small enough to ensure that the variance estimate is correct in a probabilistic sense, namely,

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

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

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

$$\kappa \le \frac{n-3}{n-1} + \left(\frac{\alpha n}{1-\alpha}\right)\left(1-\frac{1}{L^2}\right)^2 =: \kappa_{\max}(\alpha, n, L). \tag{14}$$

Figure 4 shows how large a kurtosis can be accommodated for a given n, α, and fudge factor $L = 1.5$. 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.

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. *For a given positive integer, $n_0 \in \mathbb{N}$, a fudge factor $L > 1$, and a given uncertainty tolerance, α, let $\alpha_1 = 1 - \sqrt{1-\alpha}$. Define the set of functions with bounded kurtosis:*

$$\mathcal{F}^{\text{kurt}} = \{f \in \mathcal{L}_4 : \text{kurt}(f) = \kappa \le \kappa_{\max}(n_0, \alpha_1, L)\},$$

where $\kappa_{\max}$ is defined in (14)*. For any $f \in \mathcal{F}^{\text{kurt}}$, compute the sample variance, $\hat{v}_{n_0}$ using a simple random sample of size n_0. Use this to approximate the variance of f by $\hat{\sigma}^2 = L^2\hat{v}_{n_0}$.*

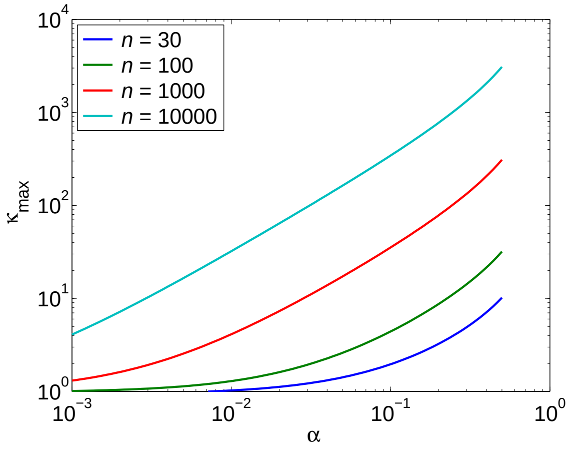

FIGURE 4. The maximum kurtosis, $\kappa_{\max}(n, \alpha, 1.5)$, as defined in (14).

Next choose an independent random sample of size

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

and compute $\hat{\mu}_n$, the simple Monte Carlo estimator of μ. Here N_C is defined in (9) *and N_B is defined in* (10b)*. A probabilistic error bound is given by*

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

Proof. By (14) it follows that $\hat{\sigma} = L\sqrt{\hat{v}_{n_0}} \ge \sigma$ with probability $1 - \alpha_1$. By (**??**) and (10), and noting that (**??**) implies that $\rho \le \kappa^{3/4}$, it follows that $\operatorname{Prob}\left[|\hat{\mu}_n - \mu| \le \epsilon\right] \ge 1 - \alpha_1$, provided that $\hat{\sigma} \ge \sigma$. Thus, the probability that both of these events happen, is at least $(1 - \alpha_1)^2 = 1 - \alpha$. □

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

$$\operatorname{cost}(\varepsilon, \sigma^2, \mathcal{F}^{\text{kurt}}, \text{aMC}) := \sup_{\substack{f \in \mathcal{F}^{\text{kurt}} \\ \operatorname{var}(f) \le \sigma^2}} \min\left\{M : \operatorname{Prob}(n_0 + n \le M) \ge 1 - \beta\right\}. \tag{15}$$

From Proposition 7 it follows that for functions in $\mathcal{F}^{\text{kurt}}$

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

where,

$$\gamma^2(\alpha_1,\beta,L) := L^2\left\{1+\sqrt{\left(\frac{\alpha_1}{1-\alpha_1}\right)\left(\frac{1-\beta}{\beta}\right)\left(1-\frac{1}{L^2}\right)^2}\right\} > 1.$$

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

$$\begin{aligned}(16)\quad \operatorname{cost}(\varepsilon,\sigma^2,\mathcal{F}^{\text{kurt}},\text{aMC}) &= \sup_{\substack{f\in\mathcal{F}^{\text{kurt}}\\ \operatorname{var}(f)\le\sigma^2}} \min\left\{M : \operatorname{Prob}(n_0+n\le M)\ge 1-\beta\right\}\\ &= n_0+\min\left\{M : \operatorname{Prob}\left(\min\left(N_C(\varepsilon/\hat\sigma,\alpha_1), N_B(\varepsilon/\hat\sigma,\alpha_1,\kappa_{\max}^{3/4})\right)\le M\right)\ge 1-\beta\right\}]]\\ &\le n_0+\min\left(N_C(\varepsilon/(\sigma\gamma(\alpha_1,\beta,L)),\alpha_1), N_B(\varepsilon/(\sigma\gamma(\alpha_1,\beta,L)),\alpha_1,\kappa_{\max}^{3/4})\right).\end{aligned}$$

Theorem 2. *The algorithm described in Theorem 1 has a probabilistic cost bounded above by* (16)*.*

The key factors that determine $\operatorname{cost}(\varepsilon,\sigma^2,\mathcal{F}^{\text{kurt}},\text{aMC})$ are ε, the error tolerance, and σ^2, the variance of the integrand. The cost is roughly proportional to $\sigma^2\varepsilon^{-2}$. For the set of integrands $\mathcal{F}^{\text{kurt}}$ the variance, $\operatorname{var}(f)$ is unbounded. Thus, for any $M>0$ there exists some $f\in\mathcal{F}^{\text{kurt}}$ for which the cost is greater than M. On the other hand, the cost does seem to behave as expected as a function of the variance of the integrand. As mentioned before, this is actually an advantage of this analysis. One need not make any assumptions about the variance of the integrand, only about the kurtosis, which is unchanged when the integrand is multiplied by an arbitrary constant.

In practice, one usually does not know, κ, the kurtosis of the integrand. The choice of n_0, L, 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, $\hat v_{n_0}$, used to determine the sample size n. Using Cantelli's inequality

$$\begin{aligned}\operatorname{Prob}(\hat v_n\ge\hat\sigma^2) &= \operatorname{Prob}(\hat v_n - L^2\hat v_{n_0}\ge 0)\\ &= \operatorname{Prob}[\hat v_n - L^2\hat v_{n_0} - (1-L^2)\sigma^2 \ge (L^2-1)\sigma^2]\\ &\le \frac{\operatorname{var}(\hat v_n - L^2\hat v_{n_0})}{\operatorname{var}(\hat v_n - L^2\hat v_{n_0}) + \{(L^2-1)\sigma^2\}^2} = \frac{1}{1+\frac{(L^2-1)^2\sigma^4}{\operatorname{var}(\hat v_n - L^2\hat v_{n_0})}}.\end{aligned}$$

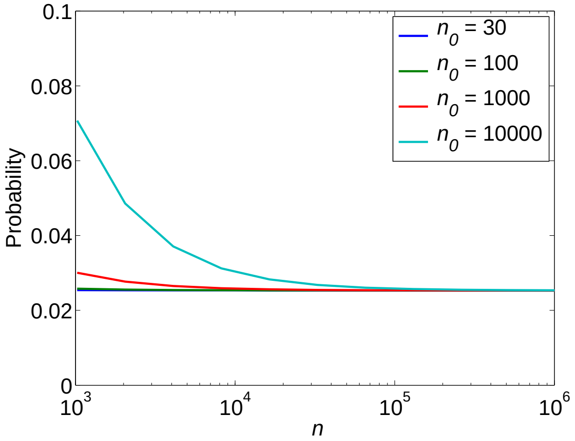

FIGURE 5. The upper bound on the probability that $\hat{v}_n \geq \hat{\sigma}^2$ in (17) for $\alpha_1 = 1 - \sqrt{95\%} \approx 2.5\%$ and $L = 1.5$.

This above quotient in the denominator can be further simplified by noticing that $\hat{v}_n$ and $\hat{v}_{n_0}$ are independent. Thus,

$$\begin{aligned}
\frac{\operatorname{var}(\hat{v}_n - L^2\hat{v}_{n_0})}{(L^2-1)^2\sigma^4} &= \frac{\operatorname{var}(\hat{v}_n) + L^4\operatorname{var}(\hat{v}_{n_0})}{(L^2-1)^2\sigma^4} \\
&= \frac{\frac{1}{n}\left(\kappa - \frac{n-3}{n-1}\right) + L^4\frac{1}{n_0}\left(\kappa - \frac{n_0-3}{n_0-1}\right)}{(L^2-1)^2} \\
&= \frac{\left(\frac{1}{n} + L^4\frac{1}{n_0}\right)\left(\kappa - \frac{n_0-3}{n_0-1}\right) + \frac{1}{n}\left(\frac{n_0-3}{n_0-1} - \frac{n-3}{n-1}\right)}{(L^2-1)^2} \\
&\leq \frac{\left(\frac{1}{n} + L^4\frac{1}{n_0}\right)\left(\frac{\alpha_1 n_0}{1-\alpha_1}\right)\left(1-\frac{1}{L^2}\right)^2 - \frac{2}{n}\left(\frac{1}{n_0-1} - \frac{1}{n-1}\right)}{(L^2-1)^2} \\
&= \left(1 + \frac{n_0}{nL^4}\right)\left(\frac{\alpha_1}{1-\alpha_1}\right) - \frac{2(n-n_0)}{n(n_0-1)(n-1)(L^2-1)^2},
\end{aligned}$$

which implies that

$$\begin{aligned}
\text{(17)} \qquad \operatorname{Prob}(\hat{v}_n \geq \hat{\sigma}^2) &\leq \frac{\left(1+\frac{n_0}{nL^4}\right)\left(\frac{\alpha_1}{1-\alpha_1}\right) - \frac{2(n-n_0)}{n(n_0-1)(n-1)(L^2-1)^2}}{1 + \left(1+\frac{n_0}{nL^4}\right)\left(\frac{\alpha_1}{1-\alpha_1}\right) - \frac{2(n-n_0)}{n(n_0-1)(n-1)(L^2-1)^2}} \\
&\leq \left(1 + \frac{n_0}{nL^4}\right)\left(\frac{\alpha_1}{1-\alpha_1}\right) \quad \text{for } n \geq n_0.
\end{aligned}$$

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 5 shows the upper bound on this probability for typical choices of α_1, L, n_0, and n.

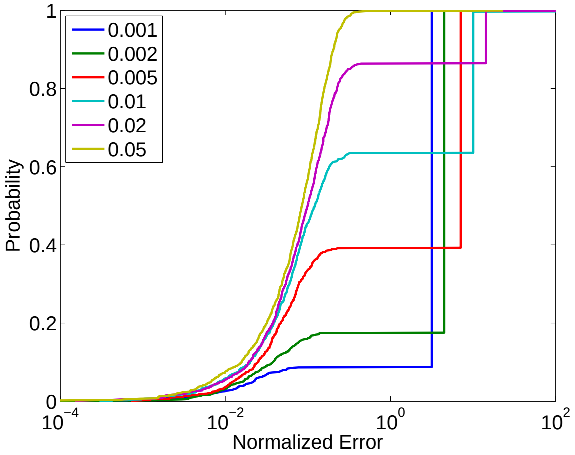

Figure 6. Empirical distribution function of $|\mu - \hat{\mu}_n|/\epsilon$ for example (18) with $\mu = \sigma = 1$, $n_0 = 100$, $\kappa_{\max} = 3.2$, $\varepsilon = 0.01$, and $p = 0.001, 0.002, 0.005, 0.01, 0.02, 0.05$ using the algorithm in Theorem 1.

5. Example

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

$$f(x) = \begin{cases} 1 + \sigma\sqrt{\frac{1-p}{p}}, & 0 \le x \le p, \\ 1 - \sigma\sqrt{\frac{p}{1-p}}, & p < x \le 1, \end{cases} \tag{18}$$

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

$$\mu = \int_0^1 g(x)\,\mathrm{d}x = 1$$

$$\operatorname{var}(g) = \int_0^1 [g(x) - \mu]^2\,\mathrm{d}x = \sigma^2 \frac{1-p}{p} p + \sigma^2 \frac{p}{1-p}(1-p) = \sigma^2,$$

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

$$= \frac{(1-p)^3 + p^3}{p(1-p)} = \frac{1 - 3p + 3p^2}{p(1-p)} = \frac{1}{p(1-p)} - 3.$$

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

Figure 6 shows the empirical distribution of the normalized error $|\mu - \hat{\mu}_n|/\epsilon$, using 1000 replications for a range of values of p. As can be seen in this figure and in Table 1, the adaptive Monte Carlo method does poorly for very small values of p, which correspond to vary large values of the kurtosis. However, even for values of the kurtosis above $\kappa_{\max} = 3.2$ used in this example, the chance of meeting the error tolerance may be quite high.

6. Questions

Here are some questions that suggest themselves:

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

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

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

Appendix of Useful Theorems

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

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

Proof. To prove Chebyshev's inequality note that

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

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

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

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

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

Then

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

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

Theorem 5. *Let $\hat{v}_n$ be the sample variance as defined in* (12)*. It's variance is*

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

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

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

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

□

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

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

Proof. Define the random variable

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

From conditional probability it is known that

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

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

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

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

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

Combining these results together yields

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

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

Proposition 7. *Let $\hat{v}_n$ be the sample variance of a function g as defined in* (12)*, and let $\kappa = \mathrm{kurt}(g)$. Then*

$$\mathrm{Prob}\left[\hat{v}_n < \sigma^2\left\{1 + \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right\}\right] \ge 1 - \alpha, \tag{19a}$$

$$\mathrm{Prob}\left[\hat{v}_n > \sigma^2\left\{1 - \sqrt{\left(\kappa - \frac{n-3}{n-1}\right)\left(\frac{1-\alpha}{\alpha n}\right)}\right\}\right] \ge 1 - \alpha. \tag{19b}$$

Proof. Choosing

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

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

$$\mathrm{Prob}[\hat{v}_n - \sigma^2 \ge a] \le \frac{\mathrm{var}(\hat{v}_n)}{a^2 + \mathrm{var}(\hat{v}_n)}$$

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

Then (19a) follows directly. By a similar argument.

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

Thus, (19b) follows as well. □

References

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

Room E1-208, Department of Applied Mathematics, Illinois Institute of Technology, 10 W. 32$^{\text{ND}}$ St., Chicago, IL 60616

Room E1-208, Department of Applied Mathematics, Illinois Institute of Technology, 10 W. 32$^{\text{ND}}$ St., Chicago, IL 60616

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