Monte Carlo with User-Specified Relative Error

J. Feng and M. Huber and Y. Ruan

Abstract Consider an estimate $\hat{a}$ for a with the property that the distribution of the relative error $\hat{a}/a - 1$ does not depend upon a, but can be chosen by the user ahead of time. Such an estimate will be said to have user-specified relative error (USRE). USRE estimates for continuous distributions such as the exponential have long been known, but only recently have unbiased USRE estimates for Bernoulli and Poisson data been discovered. In this work, biased USRE estimates are examined, and it is shown how to precisely choose the bias in order make the chance that the absolute relative error lies above a threshold decay as quickly as possible. In fact, for Poisson data this decay (on average) is slightly faster than if the CLT approximation is used.

1 Introduction

Consider the problem of generating an estimate $\hat{a}$ for a such that the relative error $(\hat{a}/a) - 1$ is bounded by user given ε, with user given failure rate δ.

Definition 1. Call an estimate $\hat{a}$ for a an (ε, δ)-*randomized approximation scheme* or (ε, δ)-*ras* for nonnegative ε and δ if

$$\mathbb{P}\left(\left|\frac{\hat{a}}{a} - 1\right| > \varepsilon\right) < \delta.$$

A stronger form is that the user actually knows precisely the distribution of the relative error.

J. Feng
Penn State University, Old Main, State College, PA, USA e-mail: jpf5265@psu.edu

M. Huber
Claremont McKenna College, 850 Columbia AV, Claremont, CA, USA e-mail: mhuber@cmc.edu

Y. Ruan
Pitzer College, 1050 N. Mills AV, Claremont, CA, USA e-mail: soruan@students.pitzer.edu

Definition 2. Say that an estimate $\hat{a}$ for a has *user-specified relative error* or *USRE* if the distribution of $\hat{a}/a$ does not depend on a, but only on parameters specified by the user in constructing $\hat{a}$.

Until recently, the only data distributions with user-specified relative error estimates were continuous and scalable.

Example 1. Say that X has an exponential distribution with rate λ (and mean $1/\lambda$) if the density of X is $f_X(s) = \lambda \exp(-\lambda s)\mathbb{1}(s \geq 0)$. Write $X \sim \mathsf{Exp}(\lambda)$. (Here $\mathbb{1}(\cdot)$ is the usual indicator function that is 1 if the argument is true and 0 if the argument is false.) Given $X_1, X_2, \ldots, X_k$ independent identically distributed (iid) data $\mathsf{Exp}(\lambda)$, an unbiased estimate for λ is

$$\hat{\lambda} = \frac{k-1}{X_1 + \cdots + X_k}.$$

Say Y has a gamma distribution with shape parameter k and rate λ (write $Y \sim \mathsf{Gamma}(k,\lambda)$) if Y has density $f_Y(s) = \lambda^k s^{k-1} \exp(-\lambda s)\mathbb{1}(s \geq 0)/\Gamma(k)$. Then it is well known that $\lambda/\hat{\lambda}$ has a gamma distribution with shape parameter k and rate parameter $k-1$. Therefore $\hat{\lambda}$ is a USRE estimate.

Example 2. Say that X is uniform over $[0,\theta]$ (write $X \sim \mathsf{Unif}([0,\theta])$) if X has density $f_X(s) = \theta^{-1}\mathbb{1}(s \in [0,\theta])$. Suppose $X_1, X_2, \ldots, X_n$ are iid $\mathsf{Unif}([0,\theta])$. Then

$$\hat{\theta} = \frac{n+1}{n} \max_i \{X_i\}$$

is an unbiased USRE estimate of θ. This is because

$$\frac{\hat{\theta}}{\theta} = \frac{n+1}{n} \max_i \left\{ \frac{X_i}{\theta} \right\},$$

and it is well known that $X_i/\theta \sim \mathsf{Unif}([0,1])$. Therefore the maximum of the X_i/θ, which is a beta distributed random variable with parameters n and 1, does not depend on $\hat{\theta}$ in any way. Such a variable has mean $n/(n+1)$, so multiplying by $(n+1)/n$ makes the estimate unbiased.

Remark 1. Throughout this work, we will always use k to denote the number of exponential random variables used in constructing our estimate. The variable n will used more generally to denote the number of samples drawn from any other distribution.

1.1 Discrete scalable distributions

The output of Monte Carlo algorithms often come from discrete rather than continuous distributions, and so the creation of user-specified relative error estimates seemed out of reach for many problems. One feature of Monte Carlo data, however, it the ability to generate as much data as needed for the estimate. That is, unlike fixed length experiments where the data output is $X_1, \ldots, X_n$, it is typically easy with Monte Carlo output to have a stream of data and use $X_1, \ldots, X_T$ for some stopping time T as the final set of data.

By carefully using this advantage and exploiting connections between discrete and continuous distributions, it was shown how to build unbiased user-specified relative error estimates for the means of Bernoulli [2] and Poisson [3] iid data.

We open here with a new estimate for the ”German tank problem”, that is, estimation of the integer θ where $X_1, X_2, \ldots$ are independent $\mathsf{Unif}(\{1,2,\ldots,\theta\})$ random variables.

Example 3. Let $X_1, X_2, \ldots$ be iid $\mathsf{Unif}(\{1,2,\ldots,\theta\})$. Then it is well known that for $U_1, U_2, \ldots$ iid $\mathsf{Unif}([0,1])$ and independent of the X_i, that $Y_i = X_i - U_i$ are iid $\mathsf{Unif}([0,\theta])$. Therefore, from Example 2, the estimate

$$\hat{\theta}_{\text{USRE}} = \frac{n+1}{n} \max_i \{X_i - U_i\}$$

is a user-specified relative error unbiased estimate of θ for the $\{X_i\}$.

The new estimate smooths the data slightly in order to obtain our USRE for θ. What do we lose by doing this? The answer is: a little, but not much. Consider the classic minimum variance unbiased estimator for θ. Given $X_1, \ldots, X_n \sim \mathsf{Unif}(\{1,2,\ldots,\theta\})$,

$$\hat{\theta}_{\text{mvue}} = \frac{1}{1+1/n} \max_i \{X_i\} - 1.$$

The variance of this estimate is

$$\mathbb{V}(\hat{\theta}_{\text{mvue}}) = \frac{(\theta - n)(\theta + 1)}{n(n+2)}.$$

Compare with the USRE, where

$$\mathbb{V}(\hat{\theta}_{\text{USRE}}) = \frac{\theta^2}{n(n+2)}$$

When $n \ll \theta$, the variances are very close together, but it always holds that the variance of the mvue is smaller than that of the USRE.

So what is lost is a small amount of variance, What is gained is the ability to give exact confidence intervals that depend very simply on the data. For instance, for $n = 35$, it holds that a beta distributed random variable with parameters n and 1 is within 10% of its maximum value with probability $1 - 0.009338$. Therefore, the same holds for $\hat{\theta}_{\text{USRE}}$, regardless of the true value of θ. Hence an exact 99% confidence interval for θ is $[\hat{\theta}_{\text{USRE}}(1-0.1), \hat{\theta}_{\text{USRE}}(1+0.1)]$.

Now consider data which is either geometric, Bernoulli, or Poisson. Table 1 gives the densities for these distributions.

Distribution	Density $f_X(s)$	Notation
Bernoulli	$p\mathbb{1}(s=1)+(1-p)\mathbb{1}(s=0)$	$\mathsf{Bern}(p)$
Geometric	$p(1-p)^{s-1}\mathbb{1}(s\in\{1,2,\ldots\})$	$\mathsf{Geo}(p)$
Poisson	$[\exp(-\mu)\mu^s/s!]\mathbb{1}(s\in\{0,1,2,\ldots\})$	$\mathsf{Pois}(\mu)$

Table 1 Discrete distributions

Example 4. Consider $G_1,G_2,\ldots,G_n \sim \mathsf{Geo}(p)$, so $\mathbb{P}(G_i=i)=p(1-p)^{i-1}$ for $i\in\{1,2,\ldots\}$. The method of moments estimator for p is

$$\hat{p}_{\text{mom}} = \frac{n}{G_1+\cdots+G_n}$$

While biased, this does converge to p with probability 1 as k goes to infinity.

As noted in [2], a USRE is obtained for geometric random variables using the following well known fact.

Lemma 1. *If $G \sim \mathsf{Geo}(p)$ and $[A|G] \sim \mathsf{Gamma}(G,1)$, then $A \sim \mathsf{Exp}(p)$.*

For each G_i, generate $[A_i|G_i] \sim \mathsf{Gamma}(G_i,1)$. By Lemma 1, each $A_i \sim \mathsf{Exp}(p)$, and then use $\hat{p}$ for p from Example 1 to obtain the USRE for p.

Example 5. For $B_1,B_2,\ldots$ iid $\mathsf{Bern}(p)$, first use the $\{B_i\}$ to generate $\{G_i\}$.

$$G_1 = \inf\{t : B_t = 1\}, \quad G_i = \inf\{t : t > G_{i-1}, B_t = 1\} - G_{i-1}.$$

Then use the $\{G_i\}$ to give $\hat{p}$ from the previous example.

Because this uses Bernoulli random variables together with gamma random variables to give the estimate, this is known as the Gamma Bernoulli Approximation Scheme (GBAS). Each geometric requires (on average) $1/p$ Bernoulli random draws to generate, so the expected number of Bernoulli random variables used by this algorithm is k/p.

The final distribution considered here, Poisson, generates a random number of exponential random variables with each Poisson by using the following well known fact about Poisson point processes.

Lemma 2. *Let $P_1,P_2,\ldots$ be iid $\mathsf{Pois}(\mu)$. Then for each interval $[i,i+1]$ for $i\in\{0,1,\ldots\}$, let C_i be a set of P_i values drawn independently and uniformly over $[i,i+1]$. Let $D_1 \leq D_2 \leq \cdots$ be the sorted values of $\cup_i C_i$. Then $D_1, D_2-D_1, D_3-D_2,\ldots$ form an iid sequence of $\mathsf{Exp}(\mu)$ random variables.*

Example 6. For $P_1,P_2,\ldots$ iid $\mathsf{Pois}(\mu)$ and fixed k, use Lemma 2 to generate $A_1,A_2,A_3,\ldots,A_k$ iid $\mathsf{Exp}(\mu)$ and then proceed as in Example 1. This estimate is called the Gamma Poisson Approximation Scheme, or GPAS for short.

Each draw of the Poisson generates (on average) μ exponential random variables, and so between k/μ and $k/\mu+1$ Poisson draws are needed (on average) to generate the exponential random variables.

1.2 Main results

Let a denote the mean of the exponential, Bernoulli, geometric, or Poisson data used to generate a random variable $R \sim \mathsf{Gamma}(k,a)$, where k is chosen by the user. Then it is simple matter to check that $\hat{a} = (k-1)/R$ is unbiased.

Since the gamma distribution is skewed, this $\hat{a}$ estimate is more likely to be too large than too small in the relative error sense. So a better estimate is

$$\hat{a}_c = \frac{k-1}{cR},$$

where c is a fixed constant. When $c = 1$, the estimate is just $\hat{a}$ which is unbiased. By choosing $c > 1$, it is possible to balance the upper and lower tails and return an estimate where the relative error is at most ε with failure probability that decays at the fastest possible rate.

The main result is the following.

Theorem 1. *Let*

$$c = \frac{2\varepsilon}{(1-\varepsilon^2)\ln(1+2\varepsilon/(1-\varepsilon))}.$$

and $\hat{a}_c = (k-1)/[cR]$ *where* $R \sim \mathsf{Gamma}(k,a)$. *Then define*

$$c_1 = \frac{1}{c(1-\varepsilon)}, \quad c_2 = \frac{1}{c(1+\varepsilon)}, \quad b(t) = te^{1-t}. \tag{1}$$

Note that $b(t) < 1$ *for* $t \neq 1$ *and for this choice of* c_1 *and* c_2, $b(c_1) = b(c_2)$, *so let* b *equal this common value. Then*

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\hat{a}_c}{a} - 1\right| > \varepsilon\right) &\leq \frac{1}{\sqrt{2\pi(k-1)}}\left[\left|\frac{c_1}{c_1-1}\right| b^{k-1} + \left|\frac{c_2}{c_2-1}\right| b^{k-1}\right] \\ &\leq \sqrt{\frac{2}{\pi\varepsilon^2(k-1)}}\exp\left(-(k-1)\left(\frac{\varepsilon^2}{2} + \frac{11\varepsilon^4}{36}\right)\right).\end{aligned}$$

By using this choice of c, it is often possible to generate an estimate with bounded relative error using fewer samples on average than a CLT analysis. For example, consider $P_1, P_2, P_3, \ldots$ iid $\mathsf{Pois}(\mu)$. The mean and variance of the $\{P_i\}$ is both μ, so consider estimating μ for $W_1, W_2, \ldots$ iid normal with mean and variance μ. The GPAS algorithm uses on average k/μ samples to generate $R \sim \mathsf{Gamma}(k,\mu)$.

So setting $n = \lfloor k/\mu \rfloor$, the sample average $\hat{\mu}_n = (W_1 + \cdots + W_n) \sim \mathsf{N}(\mu, \mu/n)$, and

$$\mathbb{P}(|(\hat{\mu}_n/\mu) - 1| > \varepsilon) > \mathbb{P}(|Z|/\sqrt{k} > \varepsilon),$$

where Z is a standard normal random variable. As shown in Section 2.1,

$$\mathbb{P}(|Z| > \varepsilon\sqrt{k}) \approx \sqrt{\frac{2}{\pi\varepsilon^2 k}}\exp\left(-k\frac{\varepsilon^2}{2}\right),$$

so when k is large, the probability for the biased Gamma concentrates slightly faster than for a normal.

For example, when $\varepsilon = 0.1$, to get $\mathbb{P}(|Z|/\sqrt{k} > \varepsilon) < 0.01$ requires $k \geq 663.4897$. But using the value of c from Theorem 1, the value of k needed using GPAS is 661. So GPAS requires on average at most $661/\mu + 1$ samples, while the normal requires at least $663/\mu$. For small μ then, the biased estimator requires fewer samples on average than the CLT approach. See Figure 1 for the failure rates as a function of k for $\varepsilon = 0.2$.

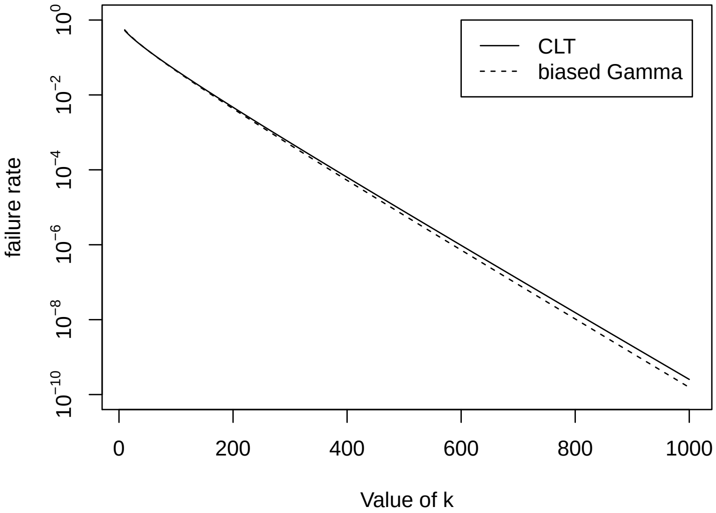

Fig. 1 Given exponential random variables with rate a, consider an estimate of a a failure if the relative error of the estimate is greater than ε. Both the problem of estimating the mean of a Bernoulli and the mean of a Poisson can be converted into this exponential problem. This plot compares the use of k exponential draws to form the estimate of a. The solid line treats the sample average of the exponentials as a normal random variable, while the dotted line uses a biased Gamma estimator. For the same k, the biased Gamma is a better estimator in this sense than the CLT. These particular failure rates use $\varepsilon = 0.2$. The CLT line has asymptotic slope against the log failure rate (to second order in ε) equal to $-\varepsilon^2/2$. The biased gamma line has asymptotic slope against the log failure rate (to the fourth order in ε) equal to $-\varepsilon^2/2 - (11/36)\varepsilon^4$.

The remainder of this work is organized as follows. The next section reviews relevant bounds on the tails of gamma and normal distributions, and proves Theorem 1. Finally, Section 3 looks at several applications of these results in Monte Carlo integration.

2 Biased estimates for minimizing the failure probability

For both GBAS and GPAS, the first step is generating a random variable $R \sim \mathsf{Gamma}(k,a)$, where a is the quantity to be estimated. Then $\hat{a}_c = (k-1)/(cR)$ becomes the estimate. The goal is to make

$$\mathbb{P}\left(\left|\frac{\hat{a}_c}{a} - 1\right| > \varepsilon\right) = \mathbb{P}\left(\frac{(k-1)}{acR} > 1+\varepsilon \text{ or } \frac{(k-1)}{acR} < 1-\varepsilon\right)$$
$$= \mathbb{P}\left(\frac{k-1}{aR} > c(1+\varepsilon)\right) + \mathbb{P}\left(\frac{k-1}{aR} < c(1-\varepsilon)\right)$$

as small as possible. Since $(aR)/(k-1) \sim \mathsf{Gamma}(k,k-1)$, our work will focus on developing good bounds for the upper and lower tails of this distribution.

Lemma 3. *Let $f_X(s) = \alpha^\beta s^{\alpha-1}\exp(-\beta s)\mathbb{1}(s \geq 0)/\Gamma(k)$ be the density of $X \sim \mathsf{Gamma}(\alpha,\beta)$. Then*

$$f_X(t)\frac{1}{\beta} \leq \mathbb{P}(X \in A) \leq f_X(t)\frac{t}{|\beta t - (\alpha-1)|}.$$

for $A = [0,t]$ where $t < (\alpha-1)/\beta$ or $A = [t,\infty)$ where $t > (\alpha-1)/\beta$.

Proof. Consider for $s > 0$,

$$f_X'(s) = f_X(s)\beta\left[\frac{\alpha-1}{\beta s} - 1\right].$$

For $s \geq t > (a-1)/\beta$, this gives

$$-\beta f_X(s) \leq f_X'(s) \leq f_X(s)\beta[(\alpha-1)/(\beta t) - 1]$$

and

$$f_X'(s)t/(\beta t - (\alpha-1)) \geq f_X(s) \geq f_X'(s)/(-\beta).$$

Integrating these inequalities for s running from t to infinity and 0 to t gives the upper and lower bounds.

The $s \leq t < (\alpha-1)/\beta$ case is similar. □

Now to understand how $f_X(s)$ behaves.

Lemma 4. *For $\alpha = k$ and $\beta = k-1$,*

$$\exp(-1/[12(k-1)])\sqrt{\frac{k-1}{2\pi}}\left(te^{1-t}\right)^{k-1} \leq f_X(t) \leq \sqrt{\frac{k-1}{2\pi}}\left(te^{1-t}\right)^{k-1}$$

Proof. Let $f_1(k-1) = \sqrt{2\pi(k-1)}((k-1)/e)^{k-1}$. Then Stirling's bound can be written

$$f_1(k-1) \leq \Gamma(k) \leq f_1(k-1)\exp(1/[12(k-1)]).$$

The density of a $\mathsf{Gamma}(k, k-1)$ at a is

$$f_X(a) = (k-1)^k t^{k-1} \exp(-(k-1)t)/\Gamma(k).$$

Using Stirling's bound on $\Gamma(k)$ and simplifying gives the result. □

Let $g(t)$ denote $\ln(\mathbb{P}((k-1)/(aR) > t))$ for $t > 1$ and $\ln(\mathbb{P}((k-1)/(aR) < t))$ for $t < 1$. From the previous lemma $g(t) = (k-1)[1-t+\ln(t)]$ plus lower order terms. Setting $w = 1-t$ gives $g(1-w) = (k-1)[w+\ln(1-w)]$. The Taylor series expansion of $g(1-w)/(k-1)$ with respect to w is

$$w + \ln(1-w) = -\frac{w^2}{2} - \frac{w^3}{3} - \frac{w^4}{4} - \cdots.$$

It is of course no surprise that the leading term of the logarithm of the tail probability is $-w^2/2$, as a $\mathsf{Gamma}(k, k-1)$ is the sum of k independent $\mathsf{Exp}(k-1)$ random variables, and therefore the CLT gives that the result is approximately normally distributed.

In the rest of this section it helps to define two values based on c and ε, as well as a function that encapsulates our rate. Recall that

$$c_1 = \frac{1}{c(1-\varepsilon)}, \quad c_2 = \frac{1}{c(1+\varepsilon)}, \quad b(t) = te^{1-t}$$

Lemma 5. *For $\hat{a}_c = (k-1)/(acR)$, let c_1, c_2, and b be as in* (1)*. Then $\mathbb{P}(|(\hat{a}_c/a) - 1| > \varepsilon)$ is in*

$$\frac{1}{\sqrt{2\pi(k-1)}}\left[b(c_1)^{k-1} + b(c_2)^{k-1}, \left|\frac{c_1}{c_1-1}\right| b(c_1)^{k-1} + \left|\frac{c_2}{c_2-1}\right| b(c_2)^{k-1}\right]$$

Proof. For $\hat{a}_c = (k-1)/(cR)$,

$$\mathbb{P}\left(\left|\frac{\hat{a}_c}{a} - 1\right| > \varepsilon\right) = \mathbb{P}\left(\frac{aR}{k-1} > \frac{1}{c(1-\varepsilon)}\right) + \mathbb{P}\left(\frac{aR}{k-1} < \frac{1}{c(1+\varepsilon)}\right)$$

Since $aR/(k-1) \sim \mathsf{Gamma}(k, k-1)$, the rest follows from the previous two lemmas. □

Since $b(t)$ is a unimodal function with maximum at $t = 1$ that goes to 0 as t goes to 0 and infinity, the log of the probability in the tail is minimized when $b(c_1) = b(c_2)$.

Lemma 6. *When*

$$c = \frac{2\varepsilon}{(1-\varepsilon^2)\ln(1+2\varepsilon/(1-\varepsilon))}, \tag{2}$$

and $\hat{a}_c = (k-1)/(cR)$, then $b(1/(c(1-\varepsilon))) = b(1/(c(1+\varepsilon))) = b$ and

$$\mathbb{P}\left(\left|\frac{\hat{a}_c}{a} - 1\right| > \varepsilon\right) \leq \frac{1}{\sqrt{2\pi(k-1)}}\left[\frac{c_1}{c_1-1} + \frac{c_2}{1-c_2}\right] b^{k-1}.$$

Proof. It is easy to verify that $b(c_1) = b(c_2)$ for this choice of c. This choice makes $c_1 > 1$ and $c_2 < 1$. Applying the previous lemma then finishes the proof. □

It helps to have an idea of how good this bound is in terms of ε. Recall that c_1, c_2, and $b = b(c_1) = b(c_2)$ are all functions of ε.

Lemma 7. *For $\varepsilon > 0$,*

$$\frac{c_1}{c_1 - 1} + \frac{c_2}{1 - c_2} \leq \frac{2}{\varepsilon}$$

and

$$b \leq \exp\left(-\frac{1}{2}\varepsilon^2 - \frac{11}{36}\varepsilon^4\right).$$

Proof. This follows directly from the Taylor series expansions of these functions in terms of ε, and the continuity of all higher derivatives for $\varepsilon > 0$. □

Combining this with the previous lemma gives the following.

Corollary 1. *For c as in* (2)*,*

$$\mathbb{P}\left(\left|\frac{k-1}{acR} - 1\right| > \varepsilon\right) \leq \sqrt{\frac{2}{\pi\varepsilon^2(k-1)}} \exp\left(-\frac{\varepsilon^2(k-1)}{2} - \frac{11\varepsilon^4(k-1)}{36}\right).$$

Therefore the log failure rate is asymptotically at most $-(k-1)(\varepsilon^2/2 + (11/36)\varepsilon^4)$. This is smaller than the asymptotic log failure rate of $-k\varepsilon^2/2$ for a normally distributed random variable.

2.1 Comparison to normal random variables

A Poisson random variable with mean μ also has variance μ. So consider $X_1, \ldots, X_n$ random variables that are normal with mean and variance μ. In Section 1 it was noted that for such random variables the sample average $\hat{\mu}_n = \sum_i X_i/n$ satisfies

$$\mathbb{P}(|(\hat{\mu}_n/\mu) - 1| > \varepsilon) = \mathbb{P}(|Z| > \varepsilon\sqrt{n\mu})$$

where Z is a standard normal random variable.

Well known bounds connect the tail probabilities of a standard normal with the density of a standard normal. For instance, Gordon [1] showed that for all $s > 0$

$$\frac{1}{s + 1/s}\frac{1}{\sqrt{2\pi}}\exp(-s^2/2) \leq \mathbb{P}(Z > s) \leq \frac{1}{s}\frac{1}{\sqrt{2\pi}}\exp(-s^2/2) \qquad (3)$$

For $s = \varepsilon\sqrt{n\mu}$, this says

$$\mathbb{P}(|\hat{\mu}_n/\mu - 1| > \varepsilon) = \Omega(\varepsilon^{-1}(n\mu)^{-1/2}\exp(-\varepsilon^2 n\mu/2)), \qquad (4)$$

(Recall that we write $f(n) = \Omega(g(n))$ if $\limsup_{n\to\infty} f(n)/g(n) > 0$.) To compare this to the failure probabilities for the Poisson random variable, note that the average number of draws of the Poisson is k/μ where k is the parameter set by the user. So if $n \approx k/\mu$, then the failure probability for the normal random variables will be

$$\Omega(\varepsilon^{-1}k^{-1/2}\exp(-\varepsilon^2 k/2),$$

while for the gamma based estimate,

$$\mathbb{P}(|\hat{p}/p-1| > \varepsilon) = O(\varepsilon^{-1}(k-1)^{-1/2}\exp(-[\varepsilon^2/2+11\varepsilon^4/36](k-1)). \quad (5)$$

So for fixed ε, as $k \to \infty$, eventually the failure probability will fall below that for the normals.

As seen in Section 1, this is not some far-off asymptotic range: for $\varepsilon = 0.1$ and $\delta = 0.01$, the gamma based method sets $k = 661$ but the normals require $k > 663$ to achieve the same level of accuracy. This fact that gammas are more highly concentrated than normals about their center is to be expected, as gamma random variables are always positive while for normals both tails are unbounded.

2.2 Biased beta estimates

Now consider the problem of estimating θ when $X_1, X_2, \ldots, X_n$ are iid $\mathsf{Unif}(\{1,2,\ldots,\theta\})$. The unbiased smoothing method generated $U_1,\ldots,U_n$ independent of $X_1,\ldots,X_n$, and set $X_i' = X_i - U_i$. This makes X_i' uniform over $[0,\theta]$. Now an unbiased USRE estimate of θ is $\hat{\theta}_{\text{USRE}} = [(n+1)/n]\max_i(X_i - U_i)$ (see Example 2.)

As earlier, given $\varepsilon > 0$, the failure probability of an estimate $\hat{\theta}$ for θ is $\mathbb{P}(|\hat{\theta}/\theta - 1| > \varepsilon)$. However, the unbiased estimate does not minimize the failure probability.

Instead, note that $\max_i(X_i - U_i) \leq \theta$, so $\hat{\theta} = (1+\varepsilon)\max_i(X_i - U_i)$ can never have relative error greater than ε. The only way the relative error can be less than $-\varepsilon$ is if $(1+\varepsilon)\max_i(X_i - U_i) < (1-\varepsilon)\theta$, or equivalently, $\max_i(X_i - U_i)/\theta < (1+\varepsilon)/(1-\varepsilon)$. Recalling that each $(X_i - U_i)/\theta \sim \mathsf{Unif}([0,1])$, this gives the following lemma.

Lemma 8. *Given $X_1,\ldots,X_n$ iid uniform over $\{1,2,\ldots,\theta\}$, and $U_1,\ldots,U_n$ iid uniform over $[0,1]$ (and independent of the $\{X_i\}$), let*

$$\hat{\theta} = (1+\varepsilon)\max_i(X_i - U_i)$$

Then

$$\mathbb{P}(|(\hat{\theta}/\theta) - 1| > \varepsilon) = \left(\frac{1-\varepsilon}{1+\varepsilon}\right)^n.$$

Since $\ln((1-\varepsilon)/(1+\varepsilon)) = -2\varepsilon - (2/3)\varepsilon^3 - \cdots$, to first order the number of samples n necessary for an (ε,δ)-ras is $(1/2)\varepsilon^{-1}\ln(\delta^{-1})$, which is very much smaller than in the exponential or normal cases.

3 Applications

This section considers applications of the GBAS and GPAS algorithms. Suppose our goal is to approximate the value of an integral of dimenson m:

$$I = \int_{x \in \mathbb{R}^m} f(x)\ \mathrm{d}x.$$

Here $f(x) \geq 0$ and m is typically very large. For instance, $f(x)$ could be the unnormalized posterior distribution of a Bayesian model (so prior density times the likelihood of data) or the solution to some #P complete problem.

Our approach is to build three sets, $C \subseteq B \subseteq A$. Set A will have Lebesgue measure equal to the integral I. Set C will have Lebesgue measure that can be computed exactly. Then, random samples will be used to estimate the ratio of the measure of A to that of B, and the ratio of the measure of B to that of C. The product then estimates that ratio of the measure of A to that of C, and then multiply by the known measure of C to estimate the measure of A which is just I.

3.1 Acceptance Rejection Integration

Using acceptance rejection to approximately integrate functions goes back to at least Von Neumann [6].

For a measure ν, say that $X \sim \nu$ over B, if for all measurable $F \subseteq B$, $\mathbb{P}(X \in F) = \nu(F)/\nu(B)$.

Given a region A, and a region B that contains A from which is possible to sample $X \sim \nu$ over B, $\mathbb{P}(X \in A) = \nu(A)/\nu(B)$. Usually it is possible to compute either $\nu(B)$ or $\nu(A)$ easily. Let $\hat{p}$ be an estimate for $\mathbb{P}(X \in A)$ obtained using biased GBAS.

If $\nu(B)$ is known, then $\hat{p}\nu(B)$ is an estimate for $\nu(A)$. If $\nu(A)$ is known then $\nu(A)/\hat{p}$ is an estimate for $\nu(A)$. Either way, to obtain $\nu(A)$ (or $\nu(B)$) within a fixed relative error requires that $\hat{p}$ estimate p within a fixed relative error.

Now consider how this idea can be turned into an algorithm for estimating I. Suppose that $f(x)$ is known through either analysis or numerical experiments to have a local maximum at x^*, and $f(x) \leq f(x^*)$ for all $x : ||x^* - x||_2 \leq \alpha$. Consider three sets,

$$\begin{aligned}
A &= \{(x,y) : x \in \mathbb{R}^n, 0 \leq y \leq f(x)\} \\
B &= \{(x,y) : ||x - x^*|| \leq \alpha, 0 \leq y \leq f(x)\} \\
C &= \{(x,y_{:}||x - x^*|| \leq \alpha, 0 \leq y \leq f(x^*)\}.
\end{aligned}$$

For ν Lebesgue measure, $\nu(A) = I$, the value of the integral that we are looking for.

It is easy to sample from C: just generate x uniformly from the hypersphere about x^* of radius α, and then generate y uniformly from $[0, f(x^*)]$.

Sampling from A is usually (approximately) accomplished using Markov chain Monte Carlo, or in some instances using perfect simulation (see [7, 4]) methods.

Then $B \subseteq A$ and $B \subseteq C$. For ν Lebesgue measure, $\nu(C) = f(x^*)\alpha^n V_n$, where V_n is the volume of an n dimensional hypersphere under $||\cdot||_2$.

So the strategy is to use two steps: estimate $\nu(B)/\nu(C)$ with $\hat{p}_1$, and $\nu(B)/\nu(A)$ with $\hat{p}_2$ using biased GBAS. Then $\nu(C)\hat{p}_1/\hat{p}_2 \approx \nu(A) = I$, and the relative error bounds for $\hat{p}_1$ and $\hat{p}_2$ can be used to find a relative error bound for the estimate of $\nu(A)$.

Of course, it is not necessary to know the value of x^* exactly. As an example, consider the function $f(x) = \exp(-x^2/2) + 1.5\exp(-(x-4)^2/2)$. Let $x^* = 0$, and $\alpha = 1$. For $x \in [-1,1]$, $f(x) \leq 1.1$. Then $A = \{(x,y) : 0 \leq y \leq f(x)\}$, $B = \{x \in [-1,1], 0 \leq y \leq f(x)\}$, and $C = \{x \in [-1,1], 0 \leq y \leq 1.1\}$. Then $\nu(C) = 2.2$, so $\nu(A) = 2.2(\nu(B)/\nu(C))/(\nu(B)/\nu(A))$.

The value of $\nu(B)/\nu(C)$ can be estimated by sampling points uniformly from C, and letting the Bernoull random variables be the indicator that the points fall into B. Similarly, the value of $\nu(B)/\nu(A)$ can be estimated by drawing samples from A and letting the Bernoulli random variables be the indcator that the points fall into B.

In this example $\nu(B)/\nu(C) \approx 0.7801$ and $\nu(B)/\nu(A) = 0.273886$. So for a given choice of k, on average $k/0.7801$ samples from C are needed to get $\hat{p}_1$ an estimate for $\nu(B)/\nu(C)$, and on average $k/0.273886$ samples from A are needed to get $\hat{p}_2$ an estimate for $\nu(B)/\nu(A)$. Recall $k = 661$ gives $\varepsilon = 0.1$ and $\delta = 0.01$. Therefore, using the union bound, $2.2\hat{p}_1/\hat{p}_2$ lies in $[(0.9/1.1)\nu(A), (1.1/0.9)\nu(A)]$ with probability at least 98%.

Suppose we use $\alpha = 0.1$. Then $\nu(B)/\nu(A) \approx 0.03187$, while $\nu(B)/\nu(C) \approx 0.908047$. The number of samples needed grows dramatically to get the $\hat{p}_2$ estimate as α becomes smaller.

3.2 TPA Integration

Generally, as α becomes smaller $\nu(B)/\nu(C)$ typically moves to 1 while $\nu(B)/\nu(A)$ becomes smaller. Therefore, it is helpful to have an alternate way to estimate $\nu(B)/\nu(A)$ when B is small relative to A. In fact, usually $\nu(B)$ is exponentially smaller than $\nu(A)$ in the dimension of the problem.

A solution to this issue is to use the Tootsie Pop Algorithm (TPA) [5]. which in this context operates as follows. Let $A_0 = A$, and draw a sample X_0 from ν over A_0. Let $A_1 = \{(x,y) : ||x - x^*|| \leq ||X_0 - x^*||\}$. Draw X_1 from ν over A_1 in the same way to get A_2, and continue into this fashion until $X_{T-1} \notin B$ and $X_T \in B$. That is, $T = \inf\{i : X_i \in B\}$.

Then Theorem 1 of [5] implies that $T - 1 \sim \textsf{Pois}(\ln(\nu(B)/\nu(A)))$. GPAS gives us an estimate $\hat{a}$ for $a = \ln(\nu(B)/\nu(A))$, along with exact confidence intervals. These in turn gives exact confidence intervals for $\exp(\hat{a})$ which estimates $\nu(B)/\nu(A)$. Combined with the exact confidence intervals for $\nu(C)/\nu(B)$, the result is an exact confidence interval for the estimate $\nu(C)\hat{p}_1\exp(-\hat{a})$ of $\nu(A)$.

Consider again our problem from earlier of estimating $\nu(B)/\nu(A)$ when the true answer is 0.0318787. Recall using $k = 661$ and directly drawing from A and forming Bernoullis from the indicator that the points fall in B used on average $k/0.0318787$ to get an estimate within relative error 0.1 with probability at least 99%.

By using TPA with $k = 661$, we obtain an estimate for $-\ln(0.0318787)$ by drawing $-661/\ln(0.0318787)$ Poisson random variables, each of which requires $-\ln(0.0318787)+1$ draws from various subsets of A. Note $(-\ln(0.0318787)+1)/(-\ln(0.0318787)) \approx 1.290$, much smaller than $1/0.0318787 \approx 31.37$.

However, the error bounds have changed. The estimate must be exponentiated to get back to the original problem. Letting $a = -\ln(0.0318787)$, we will find $\hat{a}$ such that $\hat{a} = a\xi$ where $\xi \in [0.9, 1.1]$ Hence $\exp(-a) \in [\exp(-\hat{a}/0.9), \exp(-\hat{a}/1.1)]$.

For instance, if $\hat{a} = 3.723$ (off from the true value of $a = -\ln(0.0318787) = 3.445817$) then we could say with 99% confidence that $\exp(a) = \nu(B)/\nu(A) \in [0.01597, 0.03390]$.

This is an exact confidence interval, but does not have relative error of 0.1 as desired. Using the geometric mean of the endpoints at the best estimate, the relative error could be up to 0.46. So we obtain an exact confidence interval, but not at the level of relative accuracy that we desired.

At this point, by knowing a lower bound on $\nu(B)/\nu(A)$, a second run of TPA could be undertaken that would guarantee our desired level of accuracy. Details of this two-phase procedure are given in [5].

Acknowledgements This material is based upon work supported by the National Science Foundation under Grant No. DMS 1418495. We also wish to thank the anonymous referee for several important corrections and suggestions.

References

1. Gordon, R.D.: Values of Mills' ratio of area to bounding ordinate of the normal probability integral for large values of the argument. Annals of Mathematical Statistics **12**, 364–366 (1941)
2. Huber, M.: A Bernoulli mean estimate with known relative error distribution. Random Structures Algorithms (2016). arXiv:1309.5413. To appear
3. Huber, M.: An estimator for Poisson means whose relative error distribution is known (2016). arXiv:1605.09445, Submitted
4. Huber, M.L.: Perfect Simulation. No. 148 in Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press (2015)
5. Huber, M.L., Schott, S.: Random construction of interpolating sets for high dimensional integration. Journal of Applied Probability **51**(1), 92–105 (2014). arXiv:1112.3692
6. von Neumann, J.: Various techniques used in connection with random digits. In: Monte Carlo Method, Applied Mathematics Series 12. National Bureau of Standards, Washington, D.C. (1951)
7. Propp, J.G., Wilson, D.B.: Exact sampling with coupled Markov chains and applications to statistical mechanics. Random Structures Algorithms **9**(1–2), 223–252 (1996)