Computing the Distribution Function of a Conditional Expectation via Monte Carlo: Discrete Conditioning Spaces

SHING-HOI LEE
Credit Suisse First Boston
and
PETER W. GLYNN
Stanford University

We examine different ways of numerically computing the distribution function of conditional expectations where the conditioning element takes values in a finite or countably infinite outcome space. Both the conditional expectation and the distribution function itself are computed via Monte Carlo simulation. Given a limited (and fixed) computer budget, the quality of the estimator is gauged by the inverse of its mean square error. It is a function of the fraction of the budget allocated to estimating the conditional expectation versus the amount of sampling done relative to the "conditioning variable." We will present the asymptotically optimal rates of convergence for different estimators and resolve the trade-off between the bias and variance of the estimators. Moreover, central limit theorems are established for some of the estimators proposed. We will also provide algorithms for the practical implementation of two of the estimators and illustrate how confidence intervals can be formed in each case. Major potential application areas include calculation of Value at Risk (VaR) in the field of mathematical finance and Bayesian performance analysis.

Categories and Subject Descriptors: G.3 [**Mathematics of Computing**]: Probability and Statistics

General Terms: Probability Algorithms, Distribution Functions

Additional Key Words and Phrases: Probability algorithms, distribution functions, conditional expectation

1. INTRODUCTION

Let X be a real-valued random variable (r.v.) and let Z be a random element taking values in a finite or countably infinite outcome spaces. For fixed $x \in \mathbb{R}$,

Their research was partially supported by Army Research Office Grant DAAGSS-97-0377-P0001 and National Science Foundation Grant DMS-9704732-001.
An earlier version of this article appeared in *Proceedings of Winter Simulation Conference*, 1999.
Authors' addresses: S.-H. Lee, Credit Suisse First Boston; P. W. Glynn, Management Science and Engineering, Terman Engineering Center, Room #313, Stanford University, Stanford, CA 94305-4026; email: glynn@leland.standford.edu.

our goal in this article is to compute

$$\alpha \stackrel{\triangle}{=} \mathbb{P}(\mathbb{E}\,(X|Z) \leq x). \tag{1}$$

Thus, this article is focused on computing the distribution function of a conditional expectation in the setting in which the conditioning random element Z is discrete.

To clarify the mathematical meaning of (1), it should be noted that the inner expectation involves an integration with respect to the conditional distribution of X given Z, whereas the outer probability $\mathbb{P}$ involves an integration with respect to the (unconditional) distribution of Z.

There are several different applications contexts that have served to motivate our interest in this class of problems. The first such application concerns risk management portfolios that contain substantial numbers of financial derivative options. The theory of options pricing asserts that, under suitable conditions, an option's current value can be expressed as a conditional expectation, where the conditioning random element Z is the current price of the underlying asset(s) and the expectation is computed under the so-called "equivalent martingale measure"; see, for example, Duffie [1996] for details. Consequently, if the portfolio consists of a single option, (1) expresses the probability that the value of the portfolio is less than or equal to x. More generally, if the portfolio consists of multiple options, then X is a sum over the individual replicating strategies corresponding to each of the individual options and Z is a vector corresponding to the current prices of the underlying assets.

The second major class of applications that we have in mind concerns performance evaluation problems in which statistical uncertainty exists about the dynamics of the underlying mathematical model. Assuming that the model is known up to a finite-dimensional statistical parameter, it is often appropriate to model the residual uncertainty via a posterior distribution on the parameter space that incorporates both observational data and *a priori* knowledge. When the parameter space is discrete, this can lead to a problem of the form (1) in which the conditioning random element Z is discrete. To illustrate this point, an example is in order.

Consider a telecommunications service provider that needs to make a decision regarding capacity expansion in a certain neighborhood over the next year. The goal is to deliver requested web pages to users in less than one second on average. Suppose that N is the number of subscribers in the neighborhood during the next year, and let X_i be the delivery time for the ith web page requested within the neighborhood (measured in seconds). It is reasonable to expect that $(X_i : i \geq 1)$ satisfies a law of large numbers (LLN) of the form

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}\,(X_1|N) \quad \text{almost surely.} \tag{2}$$

as $n \to \infty$, where a.s. denotes "almost surely." For example, (2) holds, under suitable conditions, if $(X_i : i \geq 1)$ is a stationary sequence which is conditionally ergodic, given N. One could attempt to design the system capacity so

that $\mathbb{E}(X|N) \leq 1$ almost surely. However, such a system design would entail building capacity appropriate to dealing with the "worst-case" performance scenario associated with the neighborhood customer base N for the provider. Substantial potential savings can be realized by instead computing $\mathbb{P}(\mathbb{E}(X|N) \leq 1)$. If this probability is sufficiently close to one, then the design capacity is deemed adequate; otherwise, it needs to be increased.

While the above examples make clear that computing (1) is of importance in certain applications, we are unaware of any existing literature dealing with this class of estimation problems. One recent relevant paper is that of Andradottir and Glynn [2002], in which the problem of computing $\alpha = \mathbb{E}\, g(\mathbb{E}(X|Z))$ is considered, where g is a differentiable function. Note that the function g appearing in (1) is an indicator function that is nonsmooth, so that the theory developed in Andradottir and Glynn [2002] is not applicable. In fact, the nonsmoothness of the indicator function has significant implications from an algorithm standpoint. While the convergence rates established in the smooth case are typically of order $c^{-1/3}$ in the computation effort of c, the corresponding estimator in the setting of this article enjoys a rate of order $(\log c/c)^{1/2}$; see Theorem 2.2. This improved convergence rate is a consequence of the fact that the bias of the estimators described in this article typically converges to zero much faster than in the smooth setting. In fact, the bias typically decays exponentially fast in c. This exponential convergence rate is a consequence of some large deviations theory that will be described more completely in Section 2, and is related to the exponential convergence theory that underlies many of the algorithms that have been proposed in connection with "ordinal optimization"; see, for example, Ho and Vakili [1992].

This article is a companion to Lee and Glynn [2002]. The latter paper explains (1) in the context of a conditioning variable Z that has a continuous distribution. The bias in the setting of continuous Z decays more slowly than for discrete Z, and this leads to a typical convergence rate of order $c^{-1/3}$ in the effort c. In other words, the presence of a continuous conditioning variable Z leads to a qualitatively different theory than that which we shall present in this article.

This article is organized as follows: Section 2 considers two estimators for α that use a fixed amount of sampling per outcome value of Z sampled to compute the conditional expectation of X given the sampled outcome value. The two estimators differ according to whether the mass function of Z is known or unknown to the simulationist. Rates of convergence are studied, and central limit theorems are obtained. Section 3 is concerned with studying the potential improvement in convergence rate that is achievable if one permits the amount of sampling done to compute $\mathbb{E}(X|Z)$ to depend on Z. The use of Z-dependent sampling to compute $\mathbb{E}(X|Z)$ improves upon the convergence rate achievable via fixed sampling only in a second-order sense. For example, when the mass function of Z is unknown, both the Z-dependent and fixed sample size estimators achieve a convergence rate of order $(\log c/c)^{1/2}$, although the constant premultiplying this factor may differ. Section 4 provides numerical results pertaining to the performance of the basic estimator, and Section 5 offers concluding remarks.

2. ESTIMATION METHODOLOGY WITH SAMPLING RATE INDEPENDENT OF OUTCOME

Suppose that the range of the random element Z is $S = \{z_1, z_2, \ldots\}$. Our discussion in this section presumes the ability of the simulationist to:

(a) draw samples from the distribution $\mathbb{P}(Z \in \cdot)$;
(b) for each z_i $(i \geq 1)$, draw samples from the conditional distribution $\mathbb{P}(X \in \cdot | Z = z_i)$.

We consider here the "obvious estimator" for α. To precisely describe this estimator, let $(Z_i : 1 \leq i \leq n)$ be a sequence of independent and identically distributed (i.i.d.) copies of the random variable Z. Conditional on $(Z_i : 1 \leq i \leq n)$, the sample $(X_j(Z_i) : 1 \leq i \leq n, 1 \leq j \leq m)$ consists of independent random variables in which $X_j(Z_i)$ follows the distribution $\mathbb{P}(X \in \cdot | Z_i)$. In other words,

$$\mathbb{P}(X_j(Z_i) \in A_{ij}, 1 \leq i \leq n, 1 \leq j \leq m | Z_i : i \geq 1) = \prod_{i=1}^{n} \prod_{j=1}^{m} \mathbb{P}(X \in A_{ij} | Z_i).$$

The obvious estimator is then

$$\alpha(m, n) = \frac{1}{n} \sum_{i=1}^{n} I(\bar{X}_m(Z_i) \leq x),$$

where $\bar{X}_m(Z_i) = m^{-1} \sum_{j=1}^{m} X_j(Z_i)$. Because the sample size m associated with $\bar{X}_m(Z_i)$ is independent of the outcome value Z_i, we call $\alpha(m, n)$ an estimator with outcome-independent sampling rate. It has expectation $\mathbb{P}(\bar{X}_m \leq x)$.

We wish to develop a central limit theorem (CLT) for this estimator that describes its rate of convergence. For a given computer budget c, let $m(c)$ and $n(c)$ be chosen so that the computational effort required to generate $\alpha(m(c), n(c))$ is approximately c. To this end, let δ_1 be the average amount of time required to generate Z_i and let δ_2 be the average amount of time required to generate $X_j(Z_i)$. Then, the aggregate effort required to compute $\alpha(m, n)$ is approximately $\delta_1 n + \delta_2 mn$. It follows that $\delta_1 n(c) + \delta_2 m(c) n(c) \approx c$. In order that $\bar{X}_{m(c)}(Z_i) \to \mathbb{E}(X | Z_i)$ almost surely as $c \to \infty$, we clearly need to impose the requirement that $m(c) \to \infty$ as $c \to \infty$. Consequently, $\delta_1 n(c) + \delta_2 m(c) n(c) \approx \delta_2 m(c) n(c)$ for c large. Finally, we may, without loss of generality, assume $\delta_2 = 1$ (for, otherwise, we can simply redefine the units by which we choose to measure computer time). Given this analysis, it is evident that $(m(c), n(c))$ must be chosen to satisfy the asymptotic relation $m(c)n(c)/c \to 1$ as $c \to \infty$.

For a given sampling plan $((m(c), n(c)) : c \geq 0)$, let $\alpha_1(c) = \alpha(m(c), n(c))$ be the estimator available after expending c units of computational time. The key to understanding the behavior of $\alpha_1(c)$ is to develop an expression for the bias of $\alpha_1(c)$. This will permit us to perform the standard "bias-variance" trade-off necessary to compute the most efficient possible sampling plan. Conditioned on

Z, we have

$$\begin{aligned}\mathbb{E}\,\alpha(m,n) &= \sum_{z\in\Gamma_+} p(z)\mathbb{P}(\bar{X}_m(z)\le x) + \sum_{z\in\Gamma_-} p(z)(1-\mathbb{P}(\bar{X}_m(z)>x)) \\ &= \alpha + \sum_{z\in\Gamma_+} p(z)\mathbb{P}(\bar{X}_m(z)\le x) - \sum_{z\in\Gamma_-} p(z)\mathbb{P}(\bar{X}_m(z)>x), \qquad (3)\end{aligned}$$

where $\Gamma_+ = \{z_i : \mathbb{E}\,(X|Z=z_i) > x, i \ge 1\}$, $\Gamma_- = \{z_i : \mathbb{E}\,(X|Z=z_i) \le x, i \ge 1\}$ and $p(z) = \mathbb{P}(Z=z)$. Thus, the rate at which the bias goes to zero is determined by the rate at which $\mathbb{P}(\bar{X}_m(z) \le x) \to 0$ for $z \in \Gamma_+$ and the rate at which $\mathbb{P}(\bar{X}_m(z) > x) \to 0$ for $z \in \Gamma_-$. These rates involving the distribution of $\bar{X}_m(z)$ are of a type that have been extensively studied as part of the substantial literature on "large deviations."

We say that a random variable X is lattice if X takes values only in the set $\{s + kd : k \in \mathbb{Z}\}$, where s is a fixed constant and d is the lattice spacing. One of the fundamental results in large deviations theory is the following (see p. 121 of Bucklew [1990]):

THEOREM 2.1. *Let $(X_i : i \ge 1)$ be an independent and identically distributed sequence of random variables such that $\mathbb{E}\,X_1 < x$. Suppose $\varphi(\theta) = \mathbb{E}\ exp(\theta X_1) < \infty$ for $\theta \in \mathbb{R}$ and that $\mathbb{P}(X_1 > x) > 0$. Then, if X_1 is not lattice, there exist finite constants η, θ^*, and σ such that*

$$\lim_{m\to\infty} \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} X_i > x\right) \exp(m\eta)\sqrt{m} = \frac{1}{\sqrt{2\pi}\sigma\theta^*}.$$

The constant θ^* is defined as the root of the equation $\varphi'(\theta^*)/\varphi(\theta^*) = x$. This root exists uniquely and the constants η and σ^2 are given by

$$\begin{aligned}\eta &= \theta^* x - \log\varphi(\theta^*), \\ \sigma^2 &= \frac{\varphi''(\theta^*)}{\varphi(\theta^*)} - x^2.\end{aligned}$$

The constants θ^*, η, and σ^2 must necessarily be strictly positive. To see this, set $\psi(\theta) = \log\varphi(\theta)$. The function $\psi(\cdot)$ is strictly convex in $[0,\infty)$ and infinitely differentiable there. Hence, $\psi''(\cdot)$ is strictly positive and $\psi'(\cdot)$ is strictly increasing on $[0,\infty)$. Observe that $\psi(0) = 0$ and $\psi'(0) = \mathbb{E}\,X_1 < x$. Because $\psi'(0) < x$, evidently the increasing nature of $\psi'(\cdot)$ implies that the root θ^* of the equation $\psi'(\theta^*) = \varphi'(\theta^*)/\varphi(\theta^*) = x$ must be strictly positive. Furthermore, the mean value theorem implies that $(\psi(\theta^*) - \psi(0))/\theta^* = \psi(\theta^*)/\theta^* = \psi'(\xi)$ for $\xi \in (0,\theta^*)$. But the increasing behavior of $\psi'(\cdot)$ implies that $\psi'(\xi) < \psi'(\theta^*) = x$. Consequently, $\psi(\theta^*)/\theta^* < x$, so that $\eta > 0$. Finally, it is easily seen that $\sigma^2 = \psi''(\theta^*)$, and hence σ^2 must be positive.

In the context of our problem, this theorem on large deviations implies that for $z \in \Gamma_-$, $\mathbb{P}(\bar{X}_m(z) > x)$ typically converges to 0 exponentially fast; whereas for $z \in \Gamma_+$, $\mathbb{P}(\bar{X}_m(z) \le x)$ typically converges to 0 exponentially fast.

More precisely, let us assume:

A1. For $i \geq 1$ and $\theta \in \mathbb{R}$, $\mathbb{E}\,[\exp(\theta X_1)|Z_1 = z_i] < \infty$;
A2. $\mathbb{P}(X_1 > x|Z_1 = z_i) > 0, z_i \in \Gamma_-$,
$\mathbb{P}(X_1 \leq x|Z_1 = z_i) > 0, z_i \in \Gamma_+$;
A3. $\mathbb{E}\,(X_1|Z_1 = z_i) \neq x$ for $i \geq 1$;
A4. For $i \geq 1$, $X_1(z_i)$ is a continuous random variable.

For $z_i \in \Gamma_-$, A1–A4 permit Theorem 2.1 to be applied directly to the random variable X, whereas for $z_i \in \Gamma_+$, A1–A4 permit Theorem 2.1 to be applied directly to $-X$. (Note, e.g., that if $z_i \in \Gamma_-$, $\mathbb{E}\,(X|Z = z_i) \leq x$. But A3 implies that $\mathbb{E}\,(X|Z = z_i) < x$). In view of Theorem 2.1, we may conclude that there exists positive constants $\eta(z_i)$ and $\gamma(z_i)$ such that

$$\mathbb{P}(\bar{X}_m(Z_1) > x|Z_1 = z_i) \sim m^{-1/2}\gamma(z_i)\exp(-m\eta(z_i)), \quad z_i \in \Gamma_- \tag{4}$$

$$\mathbb{P}(\bar{X}_m(Z_1) \leq x|Z_1 = z_i) \sim m^{-1/2}\gamma(z_i)\exp(-m\eta(z_i)), \quad z_i \in \Gamma_+ \tag{5}$$

as $m \to \infty$, where $a_m \sim b_m$ as $m \to \infty$ means that $a_m/b_m \to 1$ as $m \to \infty$. The constants $\eta(z_i)$ and $\gamma(z_i)$ appearing in (4) and (5) are given by

$$\begin{aligned}\eta(z_i) &= \theta_i^* x - \log\varphi(\theta_i^*),\\ \gamma(z_i) &= [\sqrt{2\pi}|\sigma_i\theta_i^*|]^{-1}\end{aligned}$$

where $\varphi_i(\theta) \stackrel{\triangle}{=} \mathbb{E}\,[\exp(\theta X_1)|Z_1 = z_i]$, θ_i^* is the root of $\varphi_i'(\theta_i^*)/\varphi_i(\theta_i^*) = x$, and $\sigma_i^2 = (\varphi_i''(\theta_i^*)/\varphi_i(\theta_i^*)) - x^2$.

With the aid of a couple of additional hypotheses, we can derive an asymptotic approximation for the bias of $\alpha_1(c)$. The key idea is that the bias is asymptotically determined by the values z_i for which the probabilities in (4) and (5) decay the slowest. The slowest decay rate is obviously $\eta^* = \inf\{\eta(z_i) : i \geq 1\}$.

A5. $B^* = \{z_i : i \geq 1, \eta(z_i) = \eta^*\}$ is nonempty and finite.
A6. $\inf\{\eta(z_i) : i \geq 1, z_i \notin B^*\} > \eta^*$.

Note that A5 guarantees that η^* is strictly positive, whereas A6 ensures that the exponential decay rates of the probabilities corresponding to z_i's not in B^* are strictly faster than η^*. Assumptions A5 and A6 are automatically valid, in the presence of A1–A4, when the range of Z is finite. These assumptions are discussed in greater detail in Section 5.

PROPOSITION 2.2. *Assume A1–A6. If $m(c) \to \infty$ as $c \to \infty$, then*

$$m(c)^{1/2}\exp(\eta^* m(c))(\mathbb{E}\,\alpha_1(c) - \alpha) \to \gamma^*$$

as $c \to \infty$, where

$$\gamma^* \stackrel{\triangle}{=} \sum_{z\in\Gamma_+\cap B^*} p(z)\gamma(z) - \sum_{z\in\Gamma_-\cap B^*} p(z)\gamma(z).$$

PROOF. For $\theta \in \mathbb{R}$, let $\psi(\theta; z_i) = \log(\mathbb{E}\,[\exp(\theta X_1)|Z_1 = z_i])$. Note that, for $\theta > 0$, Markov's inequality yields the bound

$$\begin{aligned}\mathbb{P}(\bar{X}_m(Z_1) > x | Z_1 = z_i) &= \mathbb{P}(\exp(\theta m \bar{X}_m(Z_1)) > \exp(\theta m x) | Z_1 = z_i) \\ &\leq \exp(-\theta m x)\mathbb{E}\left[\exp(\theta m \bar{X}_m(Z_1)) | Z_1 = z_i\right] \\ &= \exp(-\theta m x + m\psi(\theta; z_i)).\end{aligned}$$

So, if $z_i \in \Gamma_-$, we may substitute $\theta = \theta_i^*$, thereby yielding the inequality

$$\mathbb{P}(\bar{X}_m(Z_1) \leq x | Z_1 = z_i) \leq \exp(-m\eta(z_i)). \tag{6}$$

Similarly, for $\theta > 0$,

$$\begin{aligned}\mathbb{P}(\bar{X}_m(Z_1) < x | Z_1 = z_i) &= \mathbb{P}(-\bar{X}_m(Z_1) > -x | Z_1 = z_i) \\ &\leq \exp(-\theta m x + m\psi(-\theta; z_i)) \\ &= \exp(-(-\theta) m x + m\psi(-\theta; z_i)).\end{aligned}$$

For $z_i \in \Gamma_+$, $\theta_i^* < 0$. Substitute $\theta = -\theta_i^*$ into the above inequality, so that we obtain the bound

$$\mathbb{P}(\bar{X}_m(Z_1) < x | Z_1 = z_i) \leq \exp(-m\eta(z_i)). \tag{7}$$

From (3), it is evident that

$$\begin{aligned}&m(c)^{1/2}\exp(\eta^* m(c))(\mathbb{E}\,\alpha_1(c) - \alpha) \\ &= \sum_{z \in \Gamma_+ \cap B^*} p(z) m(c)^{1/2}\exp(\eta^* m(c))\mathbb{P}\left(\bar{X}_{m(c)}(z) \leq x\right) \\ &\quad - \sum_{z \in \Gamma_- \cap B^*} p(z) m(c)^{1/2}\exp(\eta^* m(c))\mathbb{P}\left(\bar{X}_{m(c)}(z) > x\right) \\ &\quad + \sum_{z \in \Gamma_+ \cap (S \setminus B^*)} p(z) m(c)^{1/2}\exp(\eta^* m(c))\mathbb{P}\left(\bar{X}_{m(c)}(z) \leq x\right) \\ &\quad - \sum_{z \in \Gamma_- \cap (S \setminus B^*)} p(z) m(c)^{1/2}\exp(\eta^* m(c))\mathbb{P}\left(\bar{X}_{m(c)}(z) > x\right).\end{aligned} \tag{8}$$

Since B^* is finite, the difference of the first two sums on the right-hand side of (8) converges to γ^*; see (4). To handle the two final sums on the right-hand side of (8), observe that (6) and (7) yield the bound

$$\begin{aligned}&\left| \sum_{z \in \Gamma_+ \cap (S \setminus B^*)} p(z) m(c)^{1/2}\exp(\eta^* m(c))\mathbb{P}\left(\bar{X}_{m(c)}(z) \leq x\right) \right. \\ &\qquad \left. - \sum_{z \in \Gamma_- \cap (S \setminus B^*)} p(z) m(c)^{1/2}\exp(\eta^* m(c))\mathbb{P}\left(\bar{X}_{m(c)}(z) > x\right) \right| \\ &\leq \sum_{z \notin B^*} p(z) m(c)^{1/2}\exp(\eta^* m(c))\exp(-\eta(z) m(c)).\end{aligned} \tag{9}$$

But according to A6, for $z \notin B^*$, $\eta(z) - \eta^*$ is uniformly positive, so

$$m(c)^{1/2}\exp(\eta^* m(c))\exp(-\eta(z) m(c)) \to 0$$

uniformly in $z \notin B^*$ as $c \to \infty$. It follows from the Dominated Convergence theorem that the right-hand side of (10) goes to zero, completing the proof. □

Hence, if $\gamma^* \neq 0$, the bias of $\alpha_1(c)$ is asymptotically equal to

$$\gamma^* m(c)^{-1/2} \exp(-\eta^* m(c)). \tag{10}$$

We now turn to the variance of $\alpha_1(c)$. Note that

$$\begin{aligned} \operatorname{Var} \alpha_1(c) &= \frac{1}{n(c)} \operatorname{Var} I\left(\bar{X}_{m(c)}(Z_1) \leq x\right) \\ &= \frac{1}{n(c)} \mathbb{E}\, \alpha_1(c)(1 - \mathbb{E}\, \alpha_1(c)) \\ &\sim \frac{1}{n(c)} \alpha(1-\alpha) \end{aligned} \tag{11}$$

as $c \to \infty$. Suppose we now choose to optimize our choice of $(m(c), n(c))$ so as to minimize

$$\operatorname{MSE}(\alpha_1(c)) = \operatorname{Var} \alpha_1(c) + (\mathbb{E}\, \alpha_1(c) - \alpha)^2$$

subject to the constraint that $m(c)n(c) \approx c$. Minimization of MSE requires roughly balancing the squared bias and the variance. In view of (10) and (11), this suggests choosing $m = m(c)$ so that

$$m \frac{\alpha(1-\alpha)}{c} \approx \frac{\gamma^{*2}}{m} \exp(-2\eta^* m).$$

Taking logarithms, we conclude that

$$-2\eta^* m - 2 \log m \approx -\log c + \log\left(\frac{\alpha(1-\alpha)}{\gamma^{*2}}\right).$$

Because $\log m$ is small relative to m (given that $m \to \infty$ is generally required for consistency), this yields the approximation

$$m \approx \frac{\log c}{2\eta^*}.$$

This asymptotic relation is supported by the following CLT for $\alpha_1(c)$; this is our main result in this section.

THEOREM 2.2. *Assume A1–A6. Suppose that $m(c) \to \infty$ and $n(c) \to \infty$ in such a way that $n(c)m(c)/c \to 1$ as $c \to \infty$. Then, if $m(c) \sim a \log c$ as $c \to \infty$ where $a \geq 1/2\eta^*$,*

$$\sqrt{\frac{c}{\log c}}\, (\alpha_1(c) - \alpha) \Rightarrow \sqrt{a\alpha(1-\alpha)}\, \mathrm{N}(0,1)$$

as $c \to \infty$, where $\Rightarrow$ denotes weak convergence and $\mathrm{N}(0,1)$ is a normally distributed random variable with mean zero and unit variance. On the other hand, if $m(c) = \lfloor a \log c \rfloor$ with $0 < a < 1/2\eta^*$, then

$$\exp(\eta^* \lfloor a \log c \rfloor) \sqrt{\log c}\, (\alpha_1(c) - \alpha) \Rightarrow \frac{\gamma^*}{\sqrt{a}}$$

as $c \to \infty$.

PROOF. We start with the case where $a \geq \frac{1}{2\eta^*}$.

Define $\chi_i(m(c)) \stackrel{\triangle}{=} I(\bar{X}_{m(c)}(Z_i) \leq x)$. Note that

$$\alpha_1(c) - \alpha = \frac{1}{n(c)} \sum_{i=1}^{n(c)} \hat{\chi}_i(m(c)) + \mathbb{P}\big(\bar{X}_{m(c)}(Z_1) \leq x\big) - \alpha,$$

where $\hat{\chi}_i(m(c)) = \chi_i(m(c)) - \mathbb{P}(\bar{X}_{m(c)}(Z) \leq x)$ is the centered version of $\chi_i(m(c))$. Then,

$$\alpha_1(c) - \alpha = [n(c)]^{-1/2} \left(\sum_{i=1}^{n(c)} \frac{\hat{\chi}_i(m(c))}{\sqrt{n(c)}} \right) + \mathbb{P}\big(\bar{X}_{m(c)}(Z) \leq x\big) - \alpha.$$

Observe that for each i,

$$|\hat{\chi}_i(m(c))| \leq 1$$

from which it follows that the family $\{\hat{\chi}_i(m(c))^2 : i = 1, \ldots, n(c), c > 0\}$ is uniformly integrable. By Lemma 1 proved in the Appendix, the Lindeberg–Feller theorem [Billingsley 1995] holds here. That is, as $c \to +\infty$,

$$\sum_{i=1}^{n(c)} \frac{\hat{\chi}_i(m(c))}{\sqrt{n(c)}} \Rightarrow \sigma \mathrm{N}(0, 1),$$

where $\sigma = \sqrt{\alpha(1-\alpha)}$.

Since $n(c)m(c)/c \to 1$ and $m(c)/\log c \to a$, we have that $c/(n(c)\log c) \to a$. Hence, by Slutzky's Theorem [Billingsley 1995], we have that

$$\sqrt{\frac{c}{\log c}} \sum_{i=1}^{n(c)} \frac{\hat{\chi}_i(m(c))}{n(c)} \Rightarrow \sqrt{a\alpha(1-\alpha)}\, \mathrm{N}(0, 1) \tag{12}$$

as $c \to \infty$. On the other hand,

$$\begin{aligned} &\sqrt{\frac{c}{\log c}} \left(\mathbb{P}\big(\bar{X}_{m(c)}(Z) \leq x\big) - \alpha \right) \\ &= \sqrt{\frac{c}{\log c}}\, m(c)^{-1/2} \exp(-\eta^* m(c)) \\ &\qquad \cdot m(c)^{1/2} \exp(\eta^* m(c)) \left(\mathbb{P}\big(\bar{X}_{m(c)}(Z) \leq x\big) - \alpha \right). \end{aligned}$$

We know that the second term converges to γ^* by Proposition 2.1. For the first term, notice that

$$\sqrt{\frac{c}{m(c)\log c}}\, \exp(-\eta^* m(c)) = c^{\frac{1}{2} - \frac{\eta^* m(c)}{\log c}} \cdot \sqrt{\frac{\log c}{m(c)}} \cdot \frac{1}{\log c}.$$

converges to 0 as $c \to \infty$ since $m(c)/\log c \to a$ and, by assumption, $1/2 - a\eta^* \leq 0$. Consequently, we must have that

$$\sqrt{\frac{c}{\log c}} \left(\mathbb{P}\big(\bar{X}_{m(c)}(Z) \leq x\big) - \alpha \right) \to 0 \tag{13}$$

as $c \to \infty$. Applying Slutzky's Theorem once again, we thus obtain the first result by combining the convergence results (12) and (13).

Similarly, if $m(c) = \lfloor a \log c \rfloor$, $m(c)n(c) \sim c$ as $c \to \infty$, and $a\eta^* < 1/2$, then, we have that

$$\exp(\eta^* \lfloor a \log c \rfloor) \sqrt{\frac{\log c}{n(c)}} \to 0$$

as $c \to \infty$. By Slutzky's Theorem,

$$\exp(\eta^* \lfloor a \log c \rfloor) \sqrt{\frac{\log c}{n(c)}} \sum_{i=1}^{n(c)} \frac{\hat{\chi}_i(m(c))}{\sqrt{n(c)}} \Rightarrow 0 \tag{14}$$

as $c \to \infty$. Also,

$$\begin{aligned} &\exp(\eta^* \lfloor a \log c \rfloor) \sqrt{\log c} \ \left(\mathbb{P}\left(\bar{X}_{m(c)}(Z) \le x \right) - \alpha \right) \\ &= \exp(\eta^* m(c)) m(c)^{1/2} \left(\mathbb{P}\left(\bar{X}_{m(c)}(Z) \le x \right) - \alpha \right) \sqrt{\frac{\log c}{m(c)}} \qquad (15) \\ &\to \frac{\gamma^*}{\sqrt{a}}. \end{aligned}$$

as $c \to \infty$. Finally, we obtain the second result by combining (14) and (16). □

In accordance with Theorem 2.2, the convergence rate of $\alpha_1(c)$ to α is roughly of order $c^{-\eta^* a}(\log c)^{-1/2}$ if $m(c) \sim a \log c$ with $a < \frac{1}{2\eta^*}$. Thus, if $a < \frac{1}{2\eta^*}$, the rate is slower than $c^{-1/2}(\log c)^{1/2}$. But Theorem 2.2 shows that the latter rate is attainable when $a \ge \frac{1}{2\eta^*}$. The variance of the limiting normal is then minimized by choosing $a = \frac{1}{2\eta^*}$. Thus, Theorem 2.2 confirms our earlier calculation which indicated that the optimal convergence rate for $\alpha_1(c)$ is attained when $m \approx (\log c)/2\eta^*$. With this choice for m, the convergence rate is within a logarithmic factor of the canonical rate of $c^{-1/2}$ that is typical of most simulation-based estimators.

However, the optimal estimator is troublesome to implement from a practical standpoint. In practice, the $\eta(z_i)$'s are difficult to estimate from the simulation output, so that η^* itself will be difficult to estimate. As a consequence, the optimal choice for m described by Theorem 2.2 may not be implementable. Furthermore, Theorem 2.2 makes clear that a poor choice of a can have a catastrophic impact in the convergence rate of $\alpha_1(c)$, particularly if a is underestimated.

In view of the potentially disastrous degradation possible if one uses a "logarithmic" sampling plan for $m(c)$ with a chosen too small, we prefer to use a more risk-averse computational strategy, in which $m(c)$ is instead permitted to grow as a power of c. The next result describes the behavior of the estimator $\alpha_1(c)$ in this setting.

THEOREM 2.3. *Assume A1–A6. Suppose that $m(c) \to \infty$ and $n(c) \to \infty$ in such a way that $m(c)n(c)/c \to 1$ as $c \to \infty$. If $m(c)/\log(c) \to \infty$, then*

$$n(c)^{1/2}(\alpha_1(c) - \alpha) \Rightarrow \sqrt{\alpha(1-\alpha)}\mathrm{N}(0,1)$$

as $c \to \infty$.

PROOF. The proof of Theorem 2.2 actually shows that if $n(c)m(c)/c \to 1$ as $c \to \infty$ with $m(c)/\log(c) \to +\infty$, and $n(c) \to +\infty$, then

$$\sqrt{n(c)}(\alpha_1(c) - \alpha) = \left(\sum_{i=1}^{n(c)} \frac{\hat{\chi}_i(m(c))}{\sqrt{n(c)}} \right) + \sqrt{n(c)} \left(\mathbb{P}\left(\bar{X}_{m(c)}(Z_1) \le x \right) - \alpha \right).$$

The first term converges in distribution to $\sqrt{\alpha(1-\alpha)}\mathrm{N}(0,1)$. The second term is given by

$$\sqrt{\frac{n(c)m(c)}{c}} \cdot c^{\frac{1}{2} - \frac{\eta^* m(c)}{\log c}} \cdot \frac{1}{m(c)} \cdot m(c)^{1/2} \exp(\eta^* m(c))(\mathbb{E}\,\alpha_1 - \alpha) \to 1 \cdot 0 \cdot 0 \cdot \gamma^* = 0.$$

Appealing to Slutzky's Theorem again, we find that

$$\sqrt{n(c)}(\alpha_1(c) - \alpha) \Rightarrow \sqrt{\alpha(1-\alpha)}\mathrm{N}(0,1)$$

as $c \to \infty$. □

It follows that, if $m(c) \sim \lfloor ac^r \rfloor$ for $a > 0$ with $r \in (0,1)$, then

$$\left[\alpha_1(c) - z\sqrt{\frac{\alpha_1(c)(1-\alpha_1(c))}{n(c)}}, \alpha_1(c) + z\sqrt{\frac{\alpha_1(c)(1-\alpha_1(c))}{n(c)}} \right]$$

is an approximate $100(1-\delta)\%$ confidence interval for α, provided c is large and z is selected so that $\mathbb{P}(-z \le \mathrm{N}(0,1) \le z) = 1 - \delta$. This suggests the following confidence interval procedure for computing α.

Algorithm 2.1

Step 0. *Initialization.* Input c, r, and a.
Step 1. *Determine the sample sizes.* Set $(m,n) \stackrel{\triangle}{=} (\lfloor ac^r \rfloor, \lfloor a^{-1}c^{1-r} \rfloor)$.
Step 2. *Determine* $\hat{\alpha}$. Set

$$\hat{\alpha} \stackrel{\triangle}{=} \frac{1}{n} \sum_{i=1}^{n} I\left(\frac{1}{m} \sum_{j=1}^{m} X_j(Z_i) \le x \right).$$

Step 3. *Form the* $100(1-\delta)\%$ *confidence interval for* $\hat{\alpha}$. Form

$$\left[\hat{\alpha} - z\sqrt{\frac{\hat{\alpha}(1-\hat{\alpha})}{n}}, \hat{\alpha} + z\sqrt{\frac{\hat{\alpha}(1-\hat{\alpha})}{n}} \right].$$

Notice that the choice of r here represents a trade-off between the bias and the variance of the estimator. In particular, for the same computation budget, the estimator has bigger bias and smaller variance when $r \ll 1$ than when r is close to 1. In Section 4, we offer empirical data associated with the performance of this confidence interval procedure for α.

We conclude this section with a discussion of an alternative estimator that is applicable when the probability mass function of Z is known. For example, in our telecommunications service provider example, it may be that the distribution of the number of subscribers is modeled via a Poisson random variable or binomial random variable, in which case the probability mass function is

known explicitly. In particular, suppose that the simulationist:

(a) has knowledge of the probability mass function $p(\cdot)$ corresponding to the random element Z;
(b) has the ability to draw samples from the conditional distribution $\mathbb{P}(X \in \cdot\,|Z = z_i)$, for each z_i $(i \geq 1)$.

The estimator we have in mind here is

$$\alpha_2(m) = \sum_i p(z_i) I(\bar{X}_m(z_i) \leq x),$$

so that the sample size used to estimate $\mathbb{P}(\mathbb{E}\,(X|Z = z_i) \leq x)$ is again outcome-independent. The computer time required to generate $\alpha_2(m)$ is proportional to m multiplied by the number of outcome values for Z. Thus, the estimator can only be (exactly) computed when the number of outcome values for Z is finite. Throughout the remainder of this section, we will assume that this is the case. Then, m scales linearly in the computer budget c so that examining the rate of convergence as a function of m is equivalent to studying the rate of convergence as a function of c.

For $i \geq 1$, let

$$\kappa_m(z_i) = \mathbb{P}(\bar{X}_m(Z_1) \leq x|Z_1 = z_i)$$

if $z_i \in \Gamma_+$ and let

$$\kappa_m(z_i) = \mathbb{P}(\bar{X}_m(Z_1) \geq x|Z_1 = z_i)$$

if $z_i \in \Gamma_-$. Then,

$$\mathbb{E}\,\alpha_2(m) - \alpha = \sum_{z \in \Gamma_+} p(z)\kappa_m(z) - \sum_{z \in \Gamma_-} p(z)\kappa_m(z)$$

and

$$\text{Var}\,\alpha_2(m) = \sum_i p(z_i)^2 \kappa_m(z_i)(1 - \kappa_m(z_i)). \tag{16}$$

Assume that A1–A4 hold (and note that A5–A6 are automatic, in view of our finite outcome assumption). It follows that Proposition 2.1 asserts that, if $\gamma^* \neq 0$, then

$$\mathbb{E}\,\alpha_2(m) - \alpha \sim \gamma^* m^{-1/2} \exp(-\eta^* m)$$

as $m \to \infty$. Furthermore, (4) and (16) together imply that

$$\text{Var}\;\alpha_2(m) \sim \beta^* m^{-1/2} \exp(-\eta^* m)$$

as $m \to \infty$, where

$$\beta^* = \sum_{z \in B^*} p(z)^2 \gamma(z).$$

As a consequence, the mean square error satisfies the asymptotic relation

$$\text{MSE}(\alpha_2(m)) \sim \beta^* m^{-1/2} \exp(-\eta^* m)$$

as $m \to \infty$, so that the mean square error converges to zero exponentially fast in this setting. Thus, in those settings where it applies, $\alpha_2(c)$ is to be preferred

to $\alpha_1(c)$, at least asymptotically (for large computer budgets). This analysis also suggests that in the large-sample context, it is the sampling of the Z-values (the "outer sampling") that contributes primarily to the variability of $\alpha_1(c)$ (rather than the "inner sampling").

3. ESTIMATION METHODOLOGY WITH OUTCOME DEPENDENT SAMPLING RATE

The large deviations asymptotics expressed by (4) assert that the impact of m on the rate at which the individual bias terms in (3) go to zero is highly state-dependent. This suggests that the amount of sampling necessary to mitigate the effect of bias is highly outcome-dependent and that improved algorithms for estimating α can, at least in principle, be obtained by permitting the "inner sample size" m to be outcome-dependent. Our goal in this section is to explore the potential increases in efficiency that can be obtained via such an idea.

We start with the case in which the mass function of Z is known. In this case, a sampling plan is an assignment of sample sizes $\vec{m} = (m(z_i) : i \geq 1)$ to each possible outcome value of Z, leading to the estimator

$$\alpha_3(\vec{m}) = \sum_i p(z_i) I\left(\bar{X}_{m(z_i)}(z_i) \leq x\right).$$

The total computer time expended to calculate $\alpha_3(\vec{m})$ is approximately proportional to $\sum_i m(z_i)$. Thus, given a computer budget c, this effectively acts as a constraint on $\sum_i m(z_i)$. We wish to find a selection of the sample sizes $(m(z_i) : i \geq 1)$ which minimizes the mean square error of $\alpha_3(\vec{m})$, subject to the constraint that $\sum_i m(z_i) \leq c$. We will denote the corresponding estimator $\alpha_3(c)$.

Assume A1–A4. To simplify the (mathematical) technical issues involved, we will suppose, through the remainder of this section, that the number of different outcome values for Z is finite, so that A5 and A6 are also in force. Just as for the estimator $\alpha_2(m)$, the bias and variance of $\alpha_3(\vec{m})$ may easily be computed:

$$\mathbb{E}\,\alpha_3(\vec{m}) - \alpha = \sum_{z \in \Gamma_+} p(z)\kappa_{m(z)}(z) - \sum_{z \in \Gamma_-} p(z)\kappa_{m(z)}(z)$$

and

$$\text{Var}\;\; \alpha_3(\vec{m}) = \sum_i p(z)^2 \kappa_{m(z_i)}(z_i)\left(1 - \kappa_{m(z_i)}(z_i)\right).$$

The following result uses the above expressions to determine the optimal sampling plan $m^* = (m_c^*(z_i) : i \geq 1)$ (for large computer budgets c) and the associated rate of convergence.

THEOREM 3.1. *Assume A1–A4 and suppose that $|S| < \infty$. Then, for any choice of $\vec{m}_c = (m_c(z_i) : i \geq 1)$ satisfying $\sum_i m_c(z_i) \leq c$,*

$$\liminf_{c \to \infty} \frac{1}{c} \log MSE(\alpha_3(\vec{m}_c)) \geq -\tau^*$$

where

$$\tau^* = \left(\sum_i \eta(z_i)^{-1}\right)^{-1}.$$

Furthermore, if we choose $m_c^*(z_i) \sim c\tau^*/\eta(z_i)$ *as* $c \to \infty$,

$$\lim_{c\to\infty} \frac{1}{c} \log MSE(\alpha_3(m_c^*)) = -\tau^*.$$

PROOF. First of all, if $m_c(z_i)$ remains bounded (in c) for some i, the mean square error does not go to zero, and the limit infimum inequality holds trivially. Hence, we assume in the remainder of the proof that $m_c(z_i) \to \infty$ for $i \geq 1$.

In this case, $x_{m_c(z_i)}(z_i) \leq 1/2$ for c large enough. So, for $\varepsilon > 0$ and c large,

$$\begin{aligned}
\text{MSE}(\alpha_3(\vec{m}_c)) &\geq \text{Var}\ \alpha_3(\vec{m}_c) \\
&= \sum_{z\in S} p(z)^2 x_{m_c(z)}(z)\big(1 - x_{m_c(z)}(z)\big) \\
&\geq \frac{1}{2}\sum_{z\in S} p(z)^2 x_{m_c(z)}(z) \\
&\geq \frac{1}{4}\sum_{z\in S} p(z)^2 \gamma(z) \exp(-(1+\varepsilon) m_c(z)\eta(z))
\end{aligned}$$

where (4) and (5) were used for the final inequality.

We now consider the minimization problem

$$\min_{\vec{m}} \sum_{z\in S} p(z)^2 \gamma(z) \exp(-(1+\varepsilon) m(z)\eta(z)) \tag{17}$$

subject to $\vec{m} = (m(z) : z \in S) \geq \vec{0}$ and $\sum_{z\in S} m(z) = c$. (Note that the continuous minimization problem has a smaller minimizer than does the problem with integer decision variables.) This optimization problem involves minimizing a strictly convex function over a convex feasible region. The minimizer may therefore be computed by the method of Lagrange multipliers. If λ is the Lagrange multiplier of the equality constraint, we find that the minimizer $m^* = (m^*(z); z \in S)$ satisfies

$$p(z)^2 \gamma(z) \exp(-(1+\varepsilon) m^*(z)\eta(z))(1+\varepsilon)\eta(z) = \lambda$$

so that

$$m^*(z) = -\frac{1}{(1+\varepsilon)\eta(z)} \log\left(\frac{\lambda}{p(z)^2\gamma(z)\eta(z)(1+\varepsilon)}\right).$$

Substituting the above expression for $m^*(z)$ into the equality constraint, we deduce that λ satisfies

$$-\log\lambda = \frac{c - \sum_{z\in S} 1/((1+\varepsilon)\eta(z)) \log(p(z)^2\gamma(z)\eta(z))}{\sum_{z\in S} 1/((1+\varepsilon)\eta(z))}.$$

So,

$$\begin{aligned}
m^*(z) &= \frac{c/\eta(z)}{\sum_{w\in S} \eta(w)^{-1}} + \frac{\log(p(z)^2)\gamma(z)\eta(z)}{(1+\varepsilon)\eta(z)} \\
&\quad - \frac{\eta(z)^{-1}}{\sum_{z\in S} \eta(w)^{-1}} \sum_{w\in S} \frac{\log(p(w)^2\gamma(w)\eta(w))}{(1+\varepsilon)\eta(w)}.
\end{aligned}$$

The minimizer of (17) is thus

$$\left(\prod_{z\in S}(p(z)^2\gamma(z)(1+\varepsilon)\eta(z))^{\tau_\varepsilon^*((1+\varepsilon)\eta(z))^{-1}}\right)\exp(-\tau_\varepsilon^* c), \tag{18}$$

where $\tau_\varepsilon^* \stackrel{\triangle}{=} (1+\varepsilon)(\sum_{z\in S}\eta(z)^{-1})^{-1}$. It follows from (18) that

$$\liminf_{c\to\infty}\frac{1}{c}\log(\mathrm{MSE}(\alpha_3(\vec{m}_c))) \geq -\tau_\varepsilon^*.$$

But $\varepsilon > 0$ was arbitrary, so we conclude that

$$\liminf_{c\to\infty}\frac{1}{c}\log(\mathrm{MSE}(\alpha_3(\vec{m}_c))) \geq -\tau^*.$$

It remains to show that τ^* is attained by the sampling plan $(m_c^*(z) : z \in S)$. We use the inequality $(\sum_{i=1}^n x_i)^2 \leq n\sum_{i=1}^n x_i^2$, as well as (4) and (5), to obtain the bound

$$\begin{aligned}\mathrm{MSE}(\alpha_3(m_c^*)) &\leq \mathrm{Var}\,\alpha_3(m_c^*) + 2|S|\sum_{z\in S}p(z)^2\frac{\gamma(z)}{m_c^*(z)}\exp(-2m_c^*(z)\eta(z))\\ &\leq (2|S|+2)\sum_{z\in S}p(z)^2\gamma(z)\exp(-m_c^*(z)\eta(z)).\end{aligned}$$

Substitution of the formula for m_c^* and straightforward analysis then establishes that

$$\limsup_{c\to\infty}\frac{1}{c}\log\mathrm{MSE}(\alpha_3(m_c^*)) \leq -\tau^*,$$

proving the result. □

It is instructive to compare the estimator $\alpha_3(m_c^*)$ to the estimator α_2 proposed in Section 2 for dealing with problems in which the mass function of Z is known. The estimator α_2 requires sampling each of Z's outcome values a constant number of times. In the notation of this section, this requires that $m(c) \approx c/|S|$. Section 2's analysis shows that $\mathrm{MSE}(\alpha_2)$ decays to zero exponentially rapidly, at a decay rate equal to $\eta^*/|S|$. It is easily seen that $\eta^* \geq \eta^*/|S|$, and hence $\alpha_3(m_c^*)$ converges at a faster exponential rate than does α_2.

Implementation of $\alpha_3(c)\,(=\alpha_3(m^*))$ requires knowledge of $\eta(z_i)$ for $i \geq 1$. Note, however, that the formula for $m_c^*(z_i)$ asserts that the most critical outcome values are those for which $\eta(z_i)$ is close to zero. In such a setting, the corresponding "large deviations" involves looking at events that are relatively more likely. Such a regime is one in which the corresponding large deviations are relatively more Gaussian (since the deviation involves a tail event that is relatively closer to the mean of the distribution). In sampling Gaussian random variables with mean $\mu < x$ and standard deviation σ, the likelihood of a deviation in the sample mean greater than x is approximately $\exp(-n(\mu-x)^2/2\sigma^2)$ (in "logarithmic scale"). This suggests the approximation $\eta(z_i) \approx (\mu(z_i)-x)^2/2\sigma^2(z_i)$ for $i \geq 1$, where $\mu(z_i)$ and $\sigma(z_i)$ are, respectively, the mean and standard deviation of the distribution $\mathbb{P}(X \in \cdot\,|Z = z_i)$. Of course, for outcome values z_i with large $\eta(z_i)$,

$(\mu(z_i) - x)^2/2\sigma^2(z_i)$ typically will not give a good approximation to its η. As mentioned earlier, such $\eta(\cdot)$'s have a small contribution to τ^*. Hence, for each z_i, this heuristic would propose spending a small portion of the computational budget to estimate $(\mu(z_i) - x)^2/2\sigma^2(z_i)$, and then using this to estimate $\eta(z_i)$, followed by "production runs" to compute α.

The algorithm below gives a practical methodology for the implementation of $\alpha_3(c)$.

Algorithm 3.1

Step 0. *Initialization.* Input $c, r \in (0, 1)$, and $\{p(z_i) : i \geq 1\}$.

Step 1. *Estimate the* $\eta(z_i)$*'s.* Let $\tilde{m} = c^r/K$. For each z_i, we sample $\tilde{m}$ X's according to $\mathbb{P}(X \in \cdot\,|Z = z_i)$ and set

$$\hat{\eta}(z_i) \stackrel{\triangle}{=} \frac{1}{2} \frac{\bar{X}_{\tilde{m}}(z_i)^2}{\frac{1}{\tilde{m}-1}\sum_{j=1}^{\tilde{m}}(X_j(z_i) - \bar{X}_{\tilde{m}}(z_i))^2}.$$

Step 2. *Estimate the optimal* $m_c^*(\cdot)$. Set

$$m^*(z_i) \stackrel{\triangle}{=} \left\lfloor \frac{\hat{\eta}(z_i)^{-1}c}{\sum_{z_i}\hat{\eta}(z_i)^{-1}} \right\rfloor.$$

Step 3. *Determine* $\hat{\alpha}$. Set

$$\alpha_3(c) = \sum_z p(z) I\left(\bar{X}_{m^*(z)}(z) \leq x\right);$$

that is, for $i \geq 1$, we sample $m^*(z_i)$ X's under the distribution function $\mathbb{P}(X \in \cdot\,|Z = z_i)$ and take $\alpha_3(c)$ as the weighted sum of the indicator functions with the weights being equal to the $p(z)$'s.

We conclude this section by discussing the use of outcome-dependent sampling in the context of random elements Z for which the probability mass function is unknown. In this setting, we must resort to sampling the Z_i's, as for the estimator $\alpha_1(c)$. Here, a sampling plan requires assigning, for a given computer budget c, an "outer sample size" $n = n(c)$. If outcome z_i is sampled, then the "inner sample size" $m = m_c(z_i)$ is utilized. For n large, the amount of "inner sampling" at outcome z_i will then be approximately $np(z_i)m_c(z_i)$ by the LLN. Consequently, the sampling plan $\vec{m} = (m_c(z_i) : i \geq 1)$ and $n = n(c)$ must be selected so that $\sum_i p(z_i)m_c(z_i) \cdot n(c) \approx c$. This leads to the estimator

$$\alpha_4(c) = \frac{1}{n(c)} \sum_{i=1}^{n(c)} I\left(\bar{X}_{m_c(Z_i)} \leq x\right).$$

Here,

$$\mathbb{E}\,\alpha_4(c) - \alpha = \sum_{z \in \Gamma_+} p(z)\kappa_{m_c(z)}(z) - \sum_{z \in \Gamma_-} p(z)\kappa_{m_c(z)}(z)$$

and

$$\text{Var}\;\alpha_4(c) = \frac{1}{n(c)}(\mathbb{E}\,\alpha_4(c))(1 - \mathbb{E}\,\alpha_4(c)).$$

An analysis very similar to that given in Section 2 for $\alpha_1(c)$ establishes the following CLT.

THEOREM 3.2. *Assume A1–A4. Suppose that for* $i \geq 1$, $m_c(z_i) \to \infty$ *and* $n(c) \to \infty$ *in such a way that* $n(c) \cdot \sum_i p(z_i)m_c(z_i)/c \to 1$ *as* $c \to \infty$. *If* $m_c(z_i) \sim a(\log c)/\eta(z_i)$ *as* $c \to \infty$ *where* $a \geq 1/2$, *then*

$$\sqrt{\frac{c}{\log c}}(\alpha_4(c) - \alpha) \Rightarrow \sqrt{a\nu\alpha(1-\alpha)}\,\mathrm{N}(0,1)$$

as $c \to \infty$, *where* $\nu = \sum_i p(z_i)/\eta(z_i)$.

Comparing Theorem 2.2 to Theorem 3.2, we see that the qualitative form of the convergence rates and limit structure is identical. Furthermore, Theorem 2.2 identifies the optimal mean-square error achievable for a given (large) value of c as approximately $(1/2\eta^*) \cdot (\log c/c) \cdot \alpha(1-\alpha)$, whereas Theorem 3.2's optimal mean-square error looks asymptotically like $(\nu/2) \cdot (\log c/c) \cdot \alpha(1-\alpha)$. Hence, the improvement obtained by using outcome-dependent sampling rates is asymptotically in proportion to $(\eta^*\nu)^{-1}$, which is always greater than or equal to one.

As for the implementation of $\alpha_4(c)$, heuristics need to be applied, in order to circumvent the difficulties inherent in $\eta(\cdot)$ being unknown. The Gaussian heuristic suggested earlier in this section is one alternative.

When compared with that of Section 2, the analysis of this section suggests that use of outcome-dependent sampling, while an improvement on outcome independent sampling, tends not to lead to a qualitatively different convergence rate. In particular, the $\sqrt{\log c/c}$ convergence rate is characteristic of both types of estimators when the mass function $p(\cdot)$ is unknown to the simulationist (and must be estimated computationally). In view of the ease of applicability of $\alpha_1(c)$, as well as its very general domain of applicability, we recommend the use of this estimator in lieu of additional problem structure that may shift the choice elsewhere.

4. NUMERICAL RESULTS

In this section, we will report on our computational experience, as obtained from a couple of examples.

Example 4.1. In this example, we assume that $Z \stackrel{\mathcal{D}}{=} \text{binomial}(10, 0.4)$ and that conditioned on $Z = z$, $X \stackrel{\mathcal{D}}{=} \mathrm{N}(z/2 - 2.3, 1)$. The exact value of $\mathbb{P}(\mathbb{E}(X|Z) \leq 0) = 0.6331032576$. We computed α using the estimator $\alpha_1(c)$.

More precisely, we implemented Algorithm 2.1 with $r = 0.2$. This value of r was chosen so that the degradation in convergence rate to order $c^{-0.4}$ was modest relative to the theoretically optimal rate of $\sqrt{\log c/c}$. The algorithm was programmed in ANSI-C and compiled and run on a Pentium II machine.

The estimator $\alpha_1(c)$ was replicated 200 independent times, at several different values of computer budget c, thereby yielding $\{\hat{\alpha}_{1,i}(c) : 1 \leq i \leq 200\}$. For each value of c, we estimated the mean, standard error, bias, and MSE of the

Table I. Numerical Results for Algorithm 2.1

c	mean	s.d.	bias	log(MSE)/c
1024	0.6454	0.1226	0.0123	−0.0041
2048	0.6475	0.0963	0.0144	−0.0023
4096	0.6454	0.0784	0.0123	−0.0012
8192	0.6386	0.0589	0.0055	−0.0007
16384	0.6383	0.0468	0.0052	−0.0004
32768	0.6347	0.0352	0.0016	−0.0002
65536	0.6332	0.0251	0.0001	−0.0001
131072	0.6332	0.0189	0.0001	−0.0001

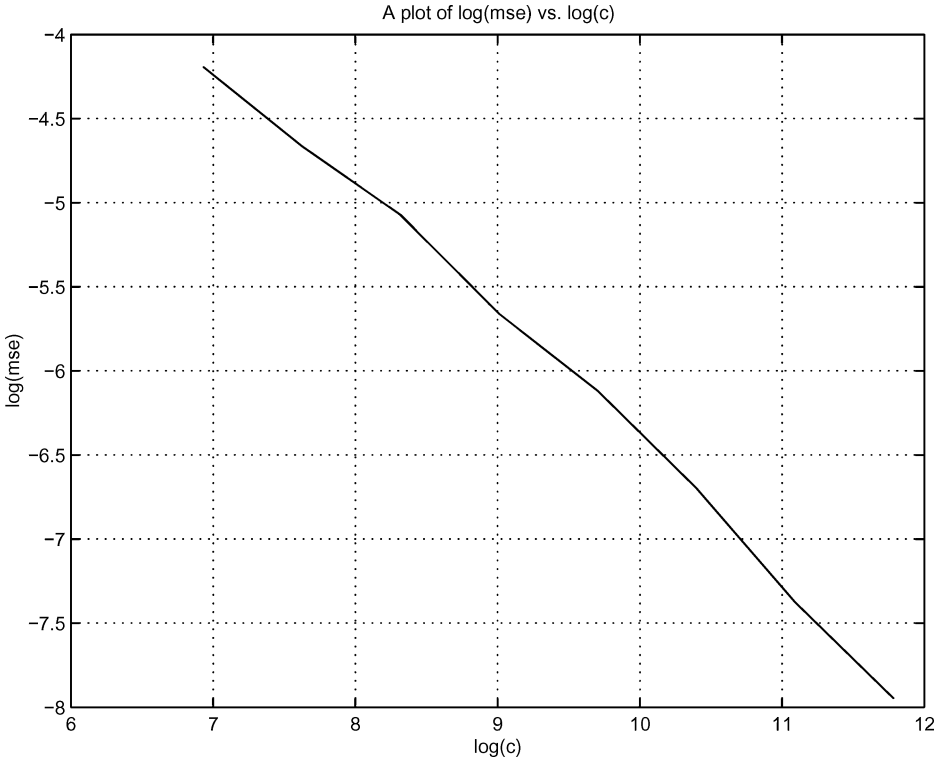

Fig. 1. Distribution function estimator for the discrete case example.

estimator as follows:

mean. set $\bar{\alpha}(c) \stackrel{\triangle}{=} (1/200)\sum_{i=1}^{200} \hat{\alpha}_{1,i}(c)$;

s.e. set $s_{\hat{\alpha}}(c) \stackrel{\triangle}{=} \sqrt{(200-1)^{-1}\sum_{i=1}^{200}(\hat{\alpha}_{1,i}(c) - \bar{\alpha}(c))^2}$;

bias. set $b_{\hat{\alpha}}(c) \stackrel{\triangle}{=} \bar{\alpha}(c) - \alpha$, where α is the exact theoretical value;

MSE. set $\mathrm{MSE}_{\hat{\alpha}}(c) \stackrel{\triangle}{=} (200)^{-1}\sum_{i=1}^{200}(\hat{\alpha}_{1,i}(c) - \alpha)^2$.

Table I summarizes the numerical results.

To deduce the rate of convergence of the estimator, we plot the log(MSE(c)) vs. $\log c$ and the plot (Figure 1) turns out to be linear. This suggests that $\mathrm{MSE}(c) \sim Vc^{\lambda}$ for some constants V and λ. We can estimate $\log V$ and λ by the y-intercept

Table II. Confidence Interval Coverage Probabilities (Std. Error in Parentheses)

c	90% cov.	95% cov.	99% cov.
1024	0.89 (0.02)	0.91 (0.02)	0.97 (0.01)
2048	0.89 (0.02)	0.93 (0.02)	0.98 (0.01)
4096	0.89 (0.02)	0.91 (0.02)	0.99 (0.01)
8192	0.91 (0.02)	0.95 (0.02)	0.98 (0.01)
16384	0.88 (0.02)	0.94 (0.02)	0.98 (0.01)
32768	0.90 (0.02)	0.95 (0.02)	0.99 (0.01)
65536	0.90 (0.02)	0.97 (0.01)	0.99 (0.01)
131072	0.91 (0.02)	0.97 (0.01)	1.00 (0.00)

Table III. Numerical Results for Estimator $\alpha_2(c)$

c	Empirical MSE	Theoretical MSE	log(MSE)/c
64	0.02189	0.02963	−0.05498
128	0.01565	0.01866	−0.03111
256	0.01014	0.01045	−0.01782
512	0.004380	0.004641	−0.01049
1024	0.001260	0.001293	−0.00649
2048	0.0001408	0.0001421	−0.00433

of the plot and its slope respectively. The theoretical slope and intercept are equal to $-(1-0.2) = -0.8$ and 0.7680 respectively; whereas the empirical slope and intercept are equal to -0.78 and 1.29 respectively, which match quite well the theoretical values.

Out of the 200 experiments, we tested the number of times, N, the confidence intervals covered the true value. The corresponding estimated coverage probability is then set to $\hat{p} \stackrel{\triangle}{=} N/200$. The standard error of the estimated coverage probability is given by $\sqrt{\hat{p}(1-\hat{p})/200}$ and is expressed inside the parenthesis beside the corresponding probability in the Table II. All coverage probabilities appear to converge to the correct values.

Example 4.2. In this example, we compute the probability in Example 4.1 using the estimator $\alpha_2(c) = \sum_i p(z_i) I(\bar{X}_m(z_i) \leq 0)$, where $m = \lfloor c/K \rfloor$ and K is the cardinality of the sample space of Z.

The estimator $\alpha_2(c)$ was replicated 200 independent times, at different values of computer budget c, thereby yielding $\{\hat{\alpha}_{2,i} : 1 \leq i \leq 200\}$. For each value of c, we estimated the MSE of the estimator as in Example 4.1.

At the end of Section 2, we showed that $\mathrm{MSE}(\alpha_2(c)) \sim \beta^*/\sqrt{c/K}\exp(-\eta^* c/K)$. In this example, $\varphi_i(\theta) = \mathbb{E}\,[\exp(\theta X)|Z = z_i] = \exp((z_i/2 - 2.3)\theta + \frac{1}{2}\theta^2)$ and $x = 0$ so that $\theta_i^* = -(z_i/2 - 2.3)$ and $\eta(z_i) = \frac{1}{2}(z_i/2 - 2.3)^2$. In particular,

$$\beta^* = \frac{\mathbb{P}(Z=5)^2}{\sqrt{2\pi}0.2} = 0.080314 \quad \text{and} \quad \eta^* = \frac{1}{2}(5/2 - 2.3)^2 = 0.02.$$

Table II summaries the numerical results. The empirical MSE's appear to converge to the correct values. Moreover, by comparing $\log(\mathrm{MSE})/c$ in the last column of Table I and Table III, we see that the MSE of α_2 converges to zero faster than that of α_1.

5. CONCLUDING REMARKS

In this article, we have offered an analysis of the behavior of several different estimators for the probability distribution of a conditional expectation, in the case that the conditioning random element Z is discrete. As pointed out in the Introduction, the behavior of such estimators when Z is continuous is quite different.

However, this article leaves several issues unanswered:

Issue 1. Assumption A1 is a condition that forces the (conditional) tails of Z to decay faster than exponentially. For example, A1 is valid for random variables that are (conditionally) bounded, Gaussian, or Weibull with shape parameter greater than one. But A1 fails for X's that are heavy tailed. For example, if X has (conditional) tails that are Pareto-like, we believe that the estimators introduced in this article will behave differently. The reason is that the bias of our estimators will no longer tend to zero exponentially in the "inner sample size" m. In particular, the rate at which the bias goes to zero can then decay at only a polynomial rate. This would change some of the analysis in this article, and is a possible topic for future research.

Issue 2. Hypotheses A5 and A6 may fail when $|S| = \infty$. This can have two consequences. The first is that the infimum η^* may be zero. The second is that the gap separating the slowest decay rate η^* from the other $\eta(z_i)$'s may be zero. If $\eta^* = 0$, the behavior of the estimators proposed here may bear some similarity to that encountered in the heavy-tailed case discussed above. If the gap vanishes, behaviors different from those described in this article are also possible. Further research on these topics may be appropriate.

Issue 3. The estimators discussed in this article generally require knowledge of the $\eta(z)$'s in order to achieve their optimal rate of convergence. Additional research is needed to develop algorithms that either explicitly or implicitly estimate the $\eta(z)$'s, with the goal of potentially achieving the same convergence rate as when the $\eta(z)$'s are known.

Nevertheless, this article develops a framework that covers many problems of potential practical interest. The limited computational experience reported here suggests that our estimators behave satisfactorily, but additional experimental study would be beneficial.

APPENDIX

LEMMA 1. *Assume that the following conditions hold:*

(*a*) *For each* $c > 0$*, the sequence* $(X_{cj} : j \geq 1)$ *consists of independent and identically distributed random variables;*
(*b*) $\mathbb{E}\, X_{c,1} = 0$, $\text{Var}\, X_{c1} \stackrel{\triangle}{=} \sigma_c^2$*;*
(*c*) $\lim_{c\to\infty} \sigma_c^2 = \sigma^2 \in (0, \infty)$*;*
(*d*) *the family* $\{X_{c,1}^2 : c > 0\}$ *is uniformly integrable.*

If $n(c) \to \infty$ *as* $c \to \infty$, *then* $\{X_{cj} : 1 \le j \le n(c), c > 0\}$ *satisfies the Lindeberg–Feller condition, namely for each* $\varepsilon > 0$,

$$\lim_{c\to\infty} \frac{1}{\text{Var } S_c} \sum_{j=1}^{n(c)} \mathbb{E}\left[X_{cj}^2 I\left(X_{cj}^2 > \varepsilon^2 \text{Var } S_c\right)\right] = 0$$

where $S_c = \sum_{j=1}^{n(c)} X_{cj}$.

PROOF. Because of conditions (a) and (b), the Lindeberg–Feller condition reduces to showing that

$$\lim_{c\to\infty} \frac{1}{\text{Var } X_{c1}} \mathbb{E}\left[X_{c1}^2 I\left(X_{c1}^2 > \varepsilon^2 n(c) \text{Var } X_{c1}\right)\right] = 0.$$

But (c) establishes that Var X_{c1} is bounded below by a positive number, so that we need only show that $\mathbb{E}\,[X_{c1}^2 I(X_{c1}^2 > \varepsilon^2 n(c)\text{Var } X_{c1})] \to 0$ as $c \to \infty$. Because $n(c)\text{Var } X_{c1} \to \infty$ as $c \to \infty$, this follows immediately from (d). □

REFERENCES

ANDRADOTTIR, S. AND GLYNN, P. 2002. Computing Bayesian means using simulation. Research Paper. S. Andradottir, School of Industrial and Systems Engineering, Georgia Institute of Technology; P. Glynn, Management Science and Engineering Dept., Stanford University.

BILLINGSLEY, P. 1995. *Probability and Measure*, 3rd ed. Wiley, New York.

BUCKLEW, J. 1990. *Large Deviation Techniques in Decision, Simulation, and Estimation*. Wiley, New York.

DUFFIE, D. 1996. *Dynamic Asset Pricing Theory*, 2nd ed. Princeton University Press, Princeton, N.J.

LEE, S. AND GLYNN, P. 2002. Computing the distribution function of a conditional expectation via Monte Carlo: Continuous conditioning spaces. Research Paper. S. Lee, Global Modeling and Analytics Group. Credit Suisse First Boston; P. Glynn, Management Science and Engineering Dept., Stanford Univerity.

Y. C. HO, R. S. AND VAKILI, P. 1992. Ordinal optimization of discrete event dynamic systems. *J. DEDS 2*, 2, 1025–1027.

Received February 2001; revised April 2003; accepted April 2003