Control Variates as Screening Functions

[Extended Abstract] *

S. Kyriazopoulou-Panagiotopoulou
Department of Informatics
Athens U of Econ & Business
Athens 10434, Greece
sofiakp@stanford.edu

I. Kontoyiannis
Department of Informatics
Athens U of Econ & Business
Athens 10434, Greece
yiannis@aueb.gr

S.P. Meyn†
Dept. of ECE & CSL
U Illinois, Urbana-Champaign
Urbana, IL 61801, USA
meyn@uiuc.edu

ABSTRACT

Suppose that the mean $\mu = E[F(X)]$ of a given function $F : \mathbb{R} \to \mathbb{R}$ is to be estimated by the empirical average $\hat{S}_n := \frac{1}{n}\sum_{i=1}^{n} F(X_i)$ of the values $F(X_i)$, where $X_1, X_2, \ldots, X_n$ are independent samples distributed like X. In cases when the mean $\nu = E[U(X)]$ of a different function $U : \mathbb{R} \to \mathbb{R}$ is known, we introduce a sampling rule, called the "screened estimator," which states that we should only consider estimates that correspond to times n when the empirical average of the $\{U(X_i)\}$ is sufficiently close to its known mean. Under the assumption that U dominates F in an appropriate sense, it is shown that the screened estimates admit exponential error bounds, even when $F(X)$ is heavy-tailed. A geometric interpretation, in the spirit of Sanov's theorem, is given for this fact, and nonasymptotic, explicit exponential bounds for the screened estimates are derived. The mathematical tools used in the analysis consist, primarily, of large deviations techniques. A detailed Markov Chain Monte Carlo (MCMC) simulation example illustrates that, in certain MCMC scenarios, screening can be very effective in terms of variance reduction, even in cases where the standard technique of control variates fails.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]; I.6 [**Simulation and Modeling**]

Keywords

Estimation, Monte Carlo, simulation, large deviations, computable bounds, measure concentration, variance reduction, Markov chain Monte Carlo (MCMC)

1. INTRODUCTION

Let $X_1, X_2, \ldots, X_n$ be independent samples distributed like a random variable X with an unknown density f on $[1, \infty)$. A common task is search for an estimator for the expectation $\mu := E[F(X)] = \int_1^\infty F(x)\, f(x)\, dx$ of some function $F : \mathbb{R} \to \mathbb{R}$ of X, and the most commonly used estimator for μ is the empirical average $\hat{S}_n$, where, for each $k \le n$, we write

$$\hat{S}_k := \frac{1}{k}\sum_{i=1}^{k} F(X_i), \quad 1 \le k \le n.$$

Let us assume that, somehow, we know two things about f: That it has a "heavy" right tail, and the value of its mean, $\nu := E(X) = \int_1^\infty x f(x)\, dx$. The tail of f is of course important for the estimation task, since relatively heavy tails imply significant variability in the data $\{X_i\}$ as well as in the subsequent estimates of μ.

For definiteness, consider a specific example where the function $F(x) = x^{3/4}$, $x \ge 1$, and the unknown density is given by $f(x) = \frac{5}{2x^{7/2}}$ for $x \ge 1$ (and $f(x) = 0$, otherwise), so that $\mu = 10/7$ and $\nu = 5/3$. Although the law of large numbers guarantees that the sequence of estimates $\{\hat{S}_k\}$ is consistent and the central limit theorem implies that the rate of convergence is of order $n^{-1/2}$, a quick glance at the behavior of $\hat{S}_k$ for finite k shows that, as expected, the estimates are highly variable: The plots in Figure 1 clearly indicate that, up to $k = n = 5000$, the $\{\hat{S}_k\}$ are still quite far from having converged. Since f is heavy tailed, this irregular behavior is hardly surprising: Indeed, the error probability $\Pr\{\hat{S}_n > \mu + \epsilon\}$ decays like,

$$\Pr\{\hat{S}_n > \mu + \epsilon\} \sim \frac{1}{\epsilon^{10/3} n^{7/3}}, \quad n \to \infty, \tag{1}$$

for any $\epsilon > 0$; see, e.g., [16]. Therefore, unlike with most classical exponential error bounds, here the error probability decays polynomially in the sample size n, and with a rather small power at that.

The main idea in this work is the proposal that the information we have about f, namely that its mean ν equals $5/3$,

*A full version of this paper is available online, at the URL: pages.cs.aueb.gr/users/yiannisk/

†Supported, in part, by grants NSF ECS-0523620 and CCF 07-29031.

ValueTools 2008, October 21 – 23, 2008, Athens, GREECE.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4242
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4242

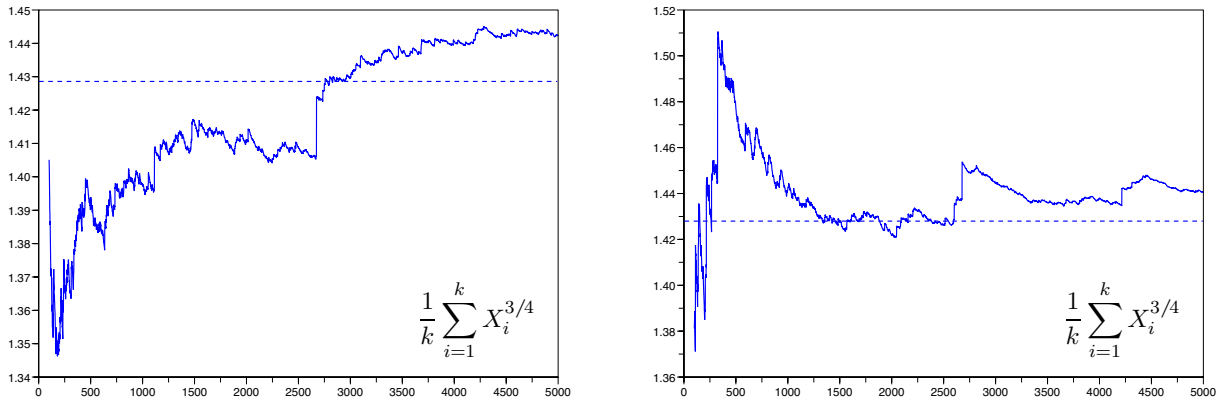

Figure 1: Two typical realizations of the estimates $\{\hat{S}_k\}$ for $k = 100, 101, \ldots, n = 5000$.

can be used to "screen" the estimates $\{\hat{S}_k\}$, as follows: Together with the $\{\hat{S}_k\}$, also compute the empirical averages $\{\hat{T}_k\}$ of the samples $\{X_i\}$ themselves,

$$\hat{T}_k = \frac{1}{k}\sum_{i=1}^{k} X_i, \quad 1 \le k \le n,$$

and *only consider estimates $\hat{S}_k$ at times k when the corresponding average $\hat{T}_k$ is within a fixed threshold $u > 0$ from its known mean.* That is, only examine $\hat{S}_k$ only at times time k such that $|\hat{T}_k - \nu| < u$. This results in what we call the "*screened estimator*" of μ. Figure 2 illustrates its performance on four different realizations of the above experiment.

More generally, assume $X, X_1, X_2, \ldots$ are independent and identically distributed (i.i.d.) random variables with unknown distribution, and we wish to estimate the expectation $\mu := E[F(X)]$ for a given function $F : \mathbb{R} \to \mathbb{R}$, while we know the value of the expectation $\nu := E[U(X)]$ of a different function $U : \mathbb{R} \to \mathbb{R}$. In this general setting, we introduce:

> **The Screened Estimator.** For each $k \ge 1$, together with the empirical averages $\{\hat{S}_k\}$ of the $\{F(X_i)\}$ also compute the averages $\{\hat{T}_k\}$ of the $\{U(X_i)\}$, and *only* consider estimates $\hat{S}_k$ at times k when $\hat{T}_k$ is within a fixed threshold $u > 0$ from its mean, i.e., those k such that, $|\hat{T}_k - \nu| < u$.

It is straightforward to see the intuition behind the above definition. In cases when we suspect that the empirical distribution $\hat{P}_k$ of the samples $\{X_i \,;\, i \le k\}$ is likely to be far from the true underlying distribution P, we can check that the projection $\int U \, d\hat{P}_k = \hat{T}_k$ of $\hat{P}_k$ along a function U is close to the projection $\int U \, dP = \nu$ of the true distribution P along U. Of course this does not guarantee that $\hat{P}_k \approx P$ or that $\hat{S}_k \approx \mu$, but it *does* rule out instances k when it is certain that $\hat{P}_k$ differs significantly from P.

Furthermore, as we shall see next, it is often possible to obtain *explicitly computable exponential error bounds* for the screened estimator, even when the error probability of the standard estimates $\{\hat{S}_k\}$ decays at a polynomial rate.

There are three main issues addressed in this work. First, in Section 2.1 we provide a theoretical explanation for the practical advantage of the screened estimator: We develop general conditions under which the error probability of the screened estimator decays exponentially, regardless of the tail of the distribution of the $\{F(X_i)\}$. The main assumption is that U dominates F from above, in that $\sup_x[F(x) - \beta U(x)]$ is finite for all $\beta > 0$, where the supremum is over all x in the support of X. Secondly, in Sections 2.2 and 2.3 we state a number of explicit exponential bounds for the error probability of the screened estimator, which are easily computable and readily applicable to specific problems where the only information we have about the unknown underlying distribution is the mean and perhaps also the variance of $U(X)$ for a particular function U. Finally, in Section 3, we present a simulation example of an estimation problem in a setting somewhat different to the setting considered so far. There, it is shown that the screened estimator can be an effective practical tool even in cases where the classical variance reduction technique of *control variates* fails; see the discussion in Sections 1.1 and 3.

In order to illustrate the effectiveness of the screened estimator, we return to the example of estimating the expectation $\mu = E(X^{3/4})$ with respect to an unknown density f on $[1, \infty)$, based on n i.i.d. samples $X_1, \ldots, X_n$ drawn from f, and assuming that we only know the mean (and perhaps some higher moments) of X. In the above notation, this corresponds to $F(x) \equiv x^{3/4}$ and $U(x) \equiv x$.

PROPOSITION 1. **(i)** The error probability of the standard estimator $\{\hat{S}_n\}$ decays to zero at a polynomial rate*: If the density f is given by $f(x) = \frac{5}{2x^{7/2}}$ for $x \ge 1$, then for any $\epsilon > 0$,*

$$\Pr\{\hat{S}_n - \mu > \epsilon\} \sim \frac{1}{\epsilon^{10/3} n^{7/3}}, \quad n \to \infty.$$

(ii) The error probability of the screened estimator decays to zero exponentially fast*: If the only information we have about f is that its mean ν equals $5/3$, then we can conclude that for all $\epsilon, u > 0$ there exists $I(\epsilon, u) > 0$ such that, for all $n \ge 1$,*

$$\Pr\{\hat{S}_n - \mu > \epsilon \;\; \textit{and} \;\; |\hat{T}_n - \tfrac{5}{3}| < u\} \le e^{-nI(\epsilon,u)}.$$

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4242
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4242

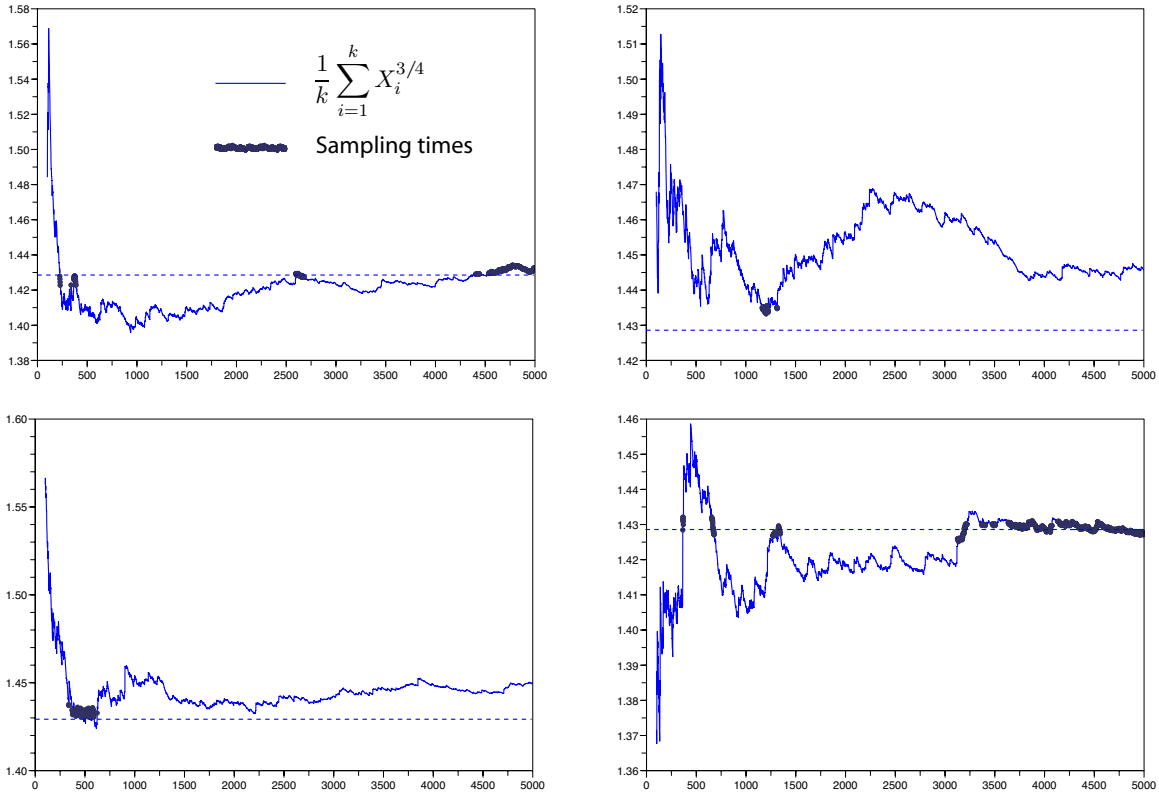

Figure 2: Four typical realizations of the estimates $\{\hat{S}_k\}$ for $k = 100, 101, \ldots, n = 5000$. The "screened estimates" are plotted in bold, and they are simply the original $\hat{S}_k$ at times k when the corresponding empirical average $\hat{T}_k$ is within $u = 0.005$ of its mean $\nu = 5/3$.

(iii) *If, in addition, we know that the variance of f equals $20/9$, then an explicit exponential bound can be computed: For any $\epsilon > 0$ and any $0 < u \le \frac{\epsilon}{20}$,*

$$\Pr\{\hat{S}_n - \mu > \epsilon \ \text{ and } \ |\hat{T}_n - \tfrac{5}{3}| < u\} \le e^{-(0.005)\times n\epsilon^2},$$

for all $n \ge 1$.

(iv) *If we also know that the value of the covariance between $X^{3/4}$ and X under f is $20/21$, then the following more accurate bound can be obtained: For any $\epsilon > 0$ and any $0 < u \le \frac{\epsilon}{20}$,*

$$\Pr\{\hat{S}_n - \mu > \epsilon \text{ and } |\hat{T}_n - \tfrac{5}{3}| < u\} \le e^{-(0.0367)\times n\epsilon^2}, \quad (2)$$

for all $n \ge 1$.

If the mean of X is known, we can employ the screened estimator and be certain that it will have an exponentially small error probability, whereas the standard estimator's probability of error may decay at least as slowly as $n^{-7/3}$. If the variance of X is also known, then for the specific values in the simulation examples in Figure 2, with $\epsilon = 0.2$, $u = 0.005$ and $n = 5000$, part (iii) of the proposition gives,

$$\Pr\{\hat{S}_n - \mu > 0.2 \ \text{ and } \ |\hat{T}_n - \tfrac{5}{3}| < 0.005\} \le 0.368.$$

This is fairly weak, despite the fact that $\epsilon = 0.2$ is a rather moderate margin of error. But the error probability does decay exponentially, and with $n = 10000$ samples the corresponding upper bound is only ≈ 0.136, while for $n = 15000$ it is ≈ 0.0498. And if, in addition, the value of the covariance between $X^{3/4}$ and X is available, then part (iv) gives a much more accurate result even for smaller ϵ: Taking $\epsilon = 0.1$, $u = 0.005$ and $n = 5000$,

$$\Pr\{\hat{S}_n - \mu > 0.1 \ \text{ and } \ |\hat{T}_n - \tfrac{5}{3}| < 0.005\} \le 0.1596,$$

and for $n = 10000$ samples the corresponding bound is ≈ 0.025.

We mention that some of the results we obtained in simulation experiments indicate that the sampling times k picked out by the screened estimator are not all equally reliable: Naturally, since the probability of error decays exponentially, earlier times correspond to much looser error bounds, while the error probability of estimates obtained during later times can be more tightly controlled. This is illustrated by the (rather atypical but not impossibly rare) results shown in Figure 3.

From the probabilistic point of view, the following calculation gives a quick explanation for the fact that the screened estimator leads to exponential error bounds in great generality (although this is not how the actual error bounds in Section 2 are obtained). Suppose the $\{\hat{S}_k\}$ are used to estimate the mean $\mu = E(F(X))$ for some F, while we know $\nu = E(U(X))$ for a different function U that dominates F in that $\operatorname{ess\,sup}_X[F(X) - \beta U(X)] < \infty$, for all $\beta > 0$. Although $F(X)$ may be heavy tailed, in which case the $\{\hat{S}_k\}$ themselves will not admit exponential error bounds, the error

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4242
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4242

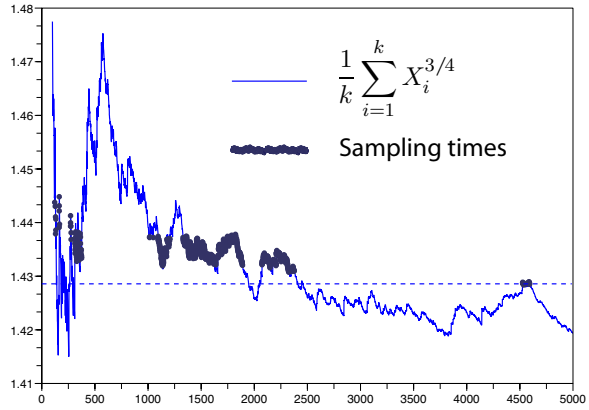

Figure 3: Another realization of the empirical estimates $\{\hat{S}_k\}$ for $k = 100, 101, \ldots, n = 5000$, plotted together with the screened estimates shown in bold (where $u = 0.005$ as before). The screened estimates at earlier times are less accurate than some of the later estimates that are ignored by the screened estimator.

probability of the screened estimator,

$$\Pr\{\hat{S}_n - \mu > \epsilon \text{ and } |\hat{T}_n - \nu| < u\},$$

is bounded above by,

$$\Pr\Big\{\frac{1}{n}\sum_{i=1}^{n}[F(X_i) - \beta U(X_i)] - (\mu - \beta\nu) > \epsilon - \beta u\Big\}. \qquad (3)$$

Since $E[F(X) - \beta U(X)] = \mu - \beta\nu$, for $0 < \beta < \frac{\epsilon}{u}$ this is a large deviations probability for the right tail of the partial sums of the random variables $\{F(X_i) - \beta U(X_i)\}$, which are (a.s.) *bounded above*. It is, therefore, no surprise that this probability is exponentially small.

1.1 Screening and control variates

A well-known and commonly used technique for reducing the variance of an estimator in classical Monte Carlo simulation is the method of *control variates*; see, e.g., the standard texts [17][14][8][1] or the paper [10] for extensive discussions. This method is based on the observation that in many applications – exactly as in our setting – there is a function U whose expectation $\nu = E[U(X)]$ is known. Therefore, replacing the estimates $\{\hat{S}_k\}$ for $\mu = E[F(X)]$ with the *control variate estimates*,

$$\tilde{S}_k := \frac{1}{k}\sum_{i=1}^{k}\Big(F(X_i) - \beta[U(X_i) - \nu]\Big), \quad 1 \le k \le n,$$

yields an estimator which is still consistent (since the additional term has zero mean) but whose variance is different from that of $\{\hat{S}_k\}$. In fact, choosing (or estimating) the value of the constant β appropriately always leads to an estimator with strictly reduced variance, as long as $F(X)$ and $U(X)$ are correlated random variables.

This technique is widely employed in practice; see the references above as well as [7][2]; also the text [9] contains many examples of current interest in computational finance and pointers to the relevant literature. In particular, functions U that appear in applications as control variates provide a natural class of screening functions that can be incorporated in the design on the screened estimator.

An interesting connection between these two methods (control variates and screening) is seen in that the probability in equation (3) above is exactly the error probability for the control variate estimates $\{\tilde{S}_k\}$.

This discussion raises an obvious question: In applications where we are not interested in actual bounds, but only in estimating μ as accurately as possible, should we use a given function U for screening or to form the standard control variate estimates $\{\tilde{S}_k\}$? In numerous simulation experiments we found that, in terms of variance reduction, screening offers no significant advantage. But we also found that in several examples of Markov Chain Monte Carlo (MCMC) estimation – the exact same setting as above, except the samples $\{X_i\}$ are generated by a Markov chain with the desired distribution as its steady-state distribution – screening was much more effective. One such example is presented in Section 3. There, screening reduces the variance of the estimates by approximately 10 to 20%, while there is *no* gain from the use of the particular function U as a control variate.

Finally, we note that no proofs are given here. For more details on these results see [13][12][4].

2. THEORY

2.1 A Geometric Explanation

In this section we give a theoretical explanation, in terms of large deviations, for the performance improvement offered by the screened estimator.

Let $X, X_1, X_2, \ldots$ be i.i.d. random variables with common law given by the probability measure P on $\mathbb{R}$. Given a function $F : \mathbb{R} \to \mathbb{R}$ whose mean is to be estimated by the empirical averages $\{\hat{S}_k\}$ of the $\{F(X_i)\}$, for the purposes of this section only we consider a slightly simplified version of the screened estimator: Assuming the mean $\nu = E(U(X))$ of a different function $U : \mathbb{R} \to \mathbb{R}$ is known, we examine the screened estimator based on the one-sided screening event, $\{\sum_{i=1}^{n} U(X_i) - n\nu < nu\}$, for some $u > 0$. To avoid cumbersome notation, write, $S_n := \sum_{i=1}^{n} F(X_i)$ and $T_n := \sum_{i=1}^{n} U(X_i)$, $n \ge 1$.

In Theorem 1 we obtain representations for the asymptotic exponents of the error probability, both for the standard estimator and for the screened estimator. The exponents are expressed in terms of relative entropy, in the spirit of Sanov's theorem; cf. [18][3][5]. Recall that the relative entropy between two probability measures P and Q on the same space is defined by,

$$H(P\|Q) := \begin{cases} \int dP \log \frac{dP}{dQ}, & \text{when } \frac{dP}{dQ} \text{ exists} \\ \infty, & \text{otherwise.} \end{cases}$$

Theorem 1 follows from the more general results in [13] and Theorem 2 below.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4242
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4242

THEOREM 1. (Sanov Asymptotics) *Suppose the functions $F : \mathbb{R} \to [0, \infty)$ and $U : \mathbb{R} \to \mathbb{R}$ have finite first moments $\mu := E[F(X)]$, $\nu := E[U(X)]$, and also finite second moments, $E[F(X)^2]$, $E[U(X)^2]$. Assume that $F(X)$ is heavy tailed in that $E[e^{\theta F(X)}] = \infty$ for all $\theta > 0$, and that U dominates F in that, $m(\beta) := \operatorname{ess\,sup}[F(X) - \beta U(X)] < \infty$ for all $\beta > 0$. Then:*

(i) The error probability of the standard estimator decays subexponentially: *For all $\epsilon > 0$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{S_n - n\mu > n\epsilon\} = -\inf_{Q\in\Sigma} H(Q\|P) = 0,$$

where Σ is the set of all probability measures Q on $\mathbb{R}$ such that $\int F dQ - \mu > \epsilon$.

(ii) The error probability of the screened estimator decays exponentially: *For all $\epsilon, u > 0$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\{S_n - n\mu > n\epsilon \text{ and } T_n - n\nu < nu\} = -\inf_{Q\in E} H(Q\|P) < 0,$$

where $E \subset \Sigma$ is the set of all probability measures Q on $\mathbb{R}$ such that $\int F dQ - \mu > \epsilon$ and $\int U dQ - \nu < u$.

Therefore, while the (asymptotic) exponent of the error probability of the standard estimator is equal to zero, the exponent of the error probability of the screened estimator is strictly positive. Although this situation is only possible when the relative entropy is minimized over an infinite-dimensional space of measures (that is, if X takes on only finitely many values, the exponent $\inf_{Q\in\Sigma} H(Q\|P)$ cannot be zero), it is perhaps illuminating to offer a geometric description.

The large oval in the first diagram in Figure 4 depicts the space of all probability measures Q on $\mathbb{R}$; the lighter areas are those Q that are "closer" to P (in that $H(Q\|P)$ is "smaller"), and the set Σ on the left consists of those Q with $\int F dQ - \mu > \epsilon$. Although P is separated from Σ by a hyperplane, we nevertheless have that $\inf_{Q\in\Sigma} H(Q\|P) = 0$. Of course this infimum is not achieved, but we can find a sequence $\{Q_n\} \subset \Sigma$, presumably near the bottom (lighter) half of Σ, such that $H(Q_n\|P) \to 0$. In the second diagram, the black shaded area corresponds to set E, formed by the intersection of Σ with the half space $H = \{Q : \int U dQ - \nu < u\}$. Note that H is a "typical" set under P, in that $P \in H$ and the empirical measure of the $\{X_i\}$ will eventually concentrate there by the ergodic theorem. Nevertheless, when Σ is intersected with H to give E, Theorem 1 tells us that it excludes the part of Σ which is close to P in relative entropy (the lighter area of Σ), and this forces the result of the minimization over $Q \in E$ to be strictly positive; the limiting minimizer Q^*, assuming it exists, is shown as laying on the common boundary of Σ and H.

The following result gives a more precise description of the large deviations *upper bounds* for the probabilities of interest. Formally, it simply establishes a version of Cramér's theorem in the present setting. What is perhaps somewhat surprising is that this is done without *any* assumptions of finite exponential moments. In the presence of the domination condition $m(\beta) < \infty$, it turns out that is only necessary to assume finite first (and in some cases second) moments for $F(X)$ and $U(X)$.

The results in Theorem 2 form the basis for the development of the bounds in Section 2.2.

THEOREM 2. (Exponential Upper Bounds) *Suppose the functions $F : \mathbb{R} \to \mathbb{R}$ and $U : \mathbb{R} \to \mathbb{R}$ are such that $\mu := E[F(X)]$ and $\nu := E[U(X)]$ are both finite, and that $m(\beta) := \operatorname{ess\,sup}[F(X) - \beta U(X)] < \infty$ for all $\beta > 0$. Then for all $\epsilon, u > 0$:*

(i) $\Pr\{S_n - n\mu > n\epsilon,\ T_n - n\nu < nu\}$ *is bounded above by $\exp\{-nH(E\|P)\}$, for all $n \geq 1$, where,*

$$H(E\|P) := \inf\{H(Q\|P) \ :\ Q \in E\}, \tag{4}$$

and E is the set of all probability measures Q on $\mathbb{R}$ such that $\int F dQ - \mu > \epsilon$ and $\int U dQ - \nu < u$.

(ii) $\Pr\{S_n - n\mu > n\epsilon,\ T_n - n\nu < nu\}$ *is bounded above by $\exp\{-n\Lambda_+^*(\epsilon, u)\}$, for all $n \geq 1$, where $\Lambda_+^*(\epsilon, u)$ is defined as*

$$\sup_{\theta_1,\theta_2\geq 0} \Big\{\theta_1(\mu+\epsilon) - \theta_2(\nu+u) - \Lambda_+(\theta_1,\theta_2)\Big\},$$

with $\Lambda_+(\theta_1, \theta_2) := \log E\Big[\exp\{\theta_1 F(X) - \theta_2 U(X)\}\Big]$, $\theta_1, \theta_2 \geq 0$.

(iii) *The rate function $\Lambda_+^*(\epsilon, u)$ is strictly positive.*

2.2 Bounds for Arbitrary Tails

Let $X, X_1, X_2, \ldots$ be i.i.d. random variables. Given functions $F, U : \mathbb{R} \to \mathbb{R}$, write $S_n = \sum_{i=1}^n F(X_i)$ and $T_n = \sum_{i=1}^n U(X_i)$. We begin by restating part of Theorem 2. Since the two-sided error event $\{S_n - n\mu > n\epsilon,\ |T_n - n\nu| < nu\}$ is contained in $\{S_n - n\mu > n\epsilon,\ \ T_n - n\nu < nu\}$, we have:

COROLLARY 1. *Suppose the functions $F : \mathbb{R} \to \mathbb{R}$ and $U : \mathbb{R} \to \mathbb{R}$ are such that $\mu := E[F(X)]$ and $\nu := E[U(X)]$ are both finite, and that $m(\beta) := \operatorname{ess\,sup}[F(X) - \beta U(X)] < \infty$ for all $\beta > 0$. Then for all $n \geq 1$ and all $\epsilon, u > 0$,*

$$\Pr\{S_n - n\mu > n\epsilon,\ \ |T_n - n\nu| < nu\} \ \leq\ e^{-n\Lambda_+^*(\epsilon,u)},$$

where the exponent, $\Lambda_+^*(\epsilon, u)$ is defined by,

$$\sup_{\theta_1\geq 0,\theta_2\geq 0} \Big\{\theta_1(\mu+\epsilon) - \theta_2(\nu+u) - \log E\Big[\exp\{\theta_1 F(X) - \theta_2 U(X)\}\Big]\Big\}, \tag{5}$$

and it is strictly positive.

If F and U also have finite second moments, an easily applicable, quantitative version of Corollary 1 can be obtained. The gist of the argument is the use of the boundedness of $[F(X) - \beta U(X)]$ in order to compute an explicit lower bound for the exponent $\Lambda_+^*(\epsilon, u)$.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4242
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4242

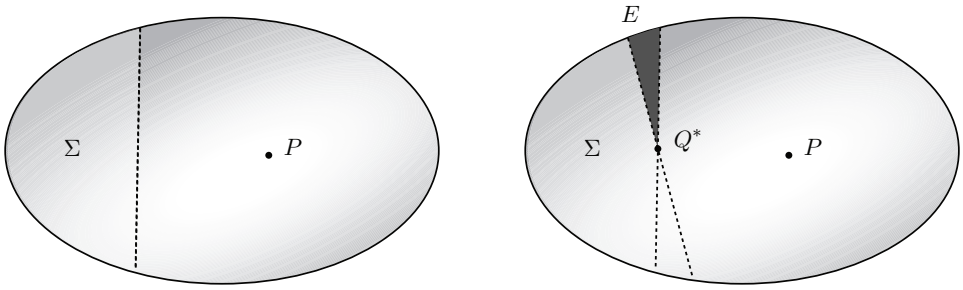

Figure 4: Geometric illustration of the fact that $\inf_{Q\in\Sigma} H(Q\|P) = 0$ whereas $\inf_{Q\in E} H(Q\|P)$ is strictly positive.

THEOREM 3. *Suppose that $E[F(X)] = E[U(X)] = 0$, that $Var(F(X)) \le 1$, $Var(U(X)) = 1$, and that $m(\beta) := \operatorname{ess\,sup}[F(X) - \beta U(X)] < \infty$ for all $\beta > 0$. Then the following hold for all $n \ge 1$:*

(i) *For any $\epsilon, u > 0$, if there exists $\beta > 0$ such that, $m(\beta) \le \epsilon - \beta u$, then,*

$$\Pr\{S_n > n\epsilon, \quad |T_n| < nu\} = 0.$$

(ii) *For any $\epsilon, u > 0$,*

$$\log \Pr\{S_n > n\epsilon, \ |T_n| < nu\} \le -2n \sup_{\alpha\in(0,1)} \left[\frac{m\cdot(1-\alpha)}{m^2+1+(\frac{\alpha\epsilon}{u})^2 - \frac{2\alpha\gamma\epsilon}{u}}\right]^2 \epsilon^2, \quad (6)$$

where $m := m(\frac{\alpha\epsilon}{u})$ and $\gamma := E[F(X)U(X)]$ is the covariance between $F(X)$ and $U(X)$.

(iii) *Let $K > 0$ arbitrary. Then for any $\epsilon > 0$ and any $0 < u \le K\epsilon$,*

$$\log \Pr\{S_n > n\epsilon, \ |T_n| < nu\} \le -\frac{n}{2}\left[\frac{M}{M^2+(1+\frac{1}{2K})^2}\right]^2 \epsilon^2, \quad (7)$$

where $M = m(\frac{1}{2K})$.

REMARKS.

1. The assumption that $\mathrm{Var}(F(X)) \le 1$ in Theorem 3 seems to require that we know an upper bound on the variance of F in advance, but in practice this is easily circumvented. In specific applications, we typically have a function U that dominates F in that, not only $m(\beta) < \infty$ for all $\beta > 0$, but also there are finite constants C_1, C_2 such that,

   $$|F(x)| \le C_1 U(x) + C_2, \ \text{ for all } x \in \text{support}(X). \quad (8)$$

   This is certainly the case for the example presented in the Introduction. A bound on the variance of $F(X)$ is obtained from (8), $\mathrm{Var}(F(X)) \le C_1^2 \mathrm{Var}(U(X)) + C_2^2$. This and several other issues arising in the application of Theorem 3 are illustrated in detail in the proof of Proposition 1 in [13].

2. In order to use the bounds in Theorem 3, it is not necessary to know $m(\beta)$ exactly; any upper bound on the $\operatorname{ess\,sup}[F(X) - \beta U(X)]$ can be used in place of $m(\beta)$. Similarly, in order to apply (6) it suffices to have an upper bound on γ, and such estimates are often easy to obtain. See the proof of Proposition 1 in [13] for an illustration.

3. The main difference between the bounds in (6) and (7) is that (7) only requires knowledge of the first and second moment of $U(X)$, whereas (6) also depends on γ. The bound in (7) is attractive because it is simple and it clearly shows that the exponent is of order ϵ^2 for small ϵ. Its main disadvantage is that it often leads to rather conservative estimates, since it ignores the potential correlation between $F(X)$ and $U(X)$ and it follows from (6) by an arbitrary choice for the parameter α. The exponent in (6), on the other hand, despite its perhaps somewhat daunting appearance, is often easy to estimate and it typically gives significantly better results. This too is clearly illustrated by the results of Proposition 1.

2.3 Bounds for Light Tails

As before, let S_n, T_n denote the partial sums of $\{F(X_i)\}$, $\{U(X_i)\}$, respectively, with respect to the i.i.d. random variables $X, X_1, X_2, \ldots$, with common law P. We assume that $E(F(X)) = E(U(X)) = 0$, and throughout this section we also assume that F and U have finite exponential moments, i.e.,

$$\Lambda(\theta) := \log E[e^{\theta F(X)}] < \infty,$$

and $E[e^{\theta U(X)}] < \infty$, for all $\theta \in \mathbb{R}$.

Corollary 1 states that the screened estimator always admits exponential error bounds, and a simple modification of its proof shows that, in fact,

$$\log \Pr\{S_n > n\epsilon, \ |T_n| < nu\} \le -n \max\{\Lambda_+^*(\epsilon, u), \Gamma_+^*(\epsilon, u)\}, \quad (9)$$

for all $n \ge 1$ and $\epsilon, u > 0$, where the exponents Λ_+^* given in (5) and

$$\Gamma_+^*(\epsilon, u) := \sup_{\theta_1\ge 0, \theta_2 \ge 0} \Big\{\theta_1(\mu+\epsilon) + \theta_2(\nu - u) - \log E\Big[\exp\{\theta_1 F(X) + \theta_2 U(X)\}\Big]\Big\},$$

are both strictly positive. But in this setting, the standard estimates $\hat{S}_n = \frac{1}{n} S_n$ also admit exponential error bounds; Cramér's theorem states that,

$$\log \Pr\{S_n > n\epsilon\} \le -n\Lambda^*(\epsilon), \quad n \ge 1, \quad (10)$$

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4242
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4242

where

$$\Lambda^*(\epsilon) := \sup_{\theta \geq 0}\{\theta\epsilon - \Lambda(\theta)\} > 0,$$

for any $\epsilon > 0$; cf., [5]. Note that the exponents in both (9) and (10) are asymptotically tight.

In this section we develop conditions under which it can be shown that the screened estimator offers a nontrivial improvement. That is, even when the error of the standard estimator decays exponentially, the error of the screened estimator has a better rate in the exponent. To that end, we look at difference,

$$\Delta(\epsilon, u) := \max\{\Lambda^*_+(\epsilon, u), \Gamma^*_+(\epsilon, u)\} - \Lambda^*(\epsilon).$$

Clearly $\Delta(\epsilon, u)$ is always nonnegative. Theorem 4 says that, as long as the covariance between $F(X)$ and $U(X)$ is not zero, $\Delta(\epsilon, u)$ is strictly positive for all ϵ, u small enough. This is strengthened in Theorem 5, where it is shown that this improvement is a "first order effect," in that, for small ϵ, u, $\Delta(\epsilon, u)$ and $\max\{\Lambda^*_+(\epsilon, u), \Gamma^*_+(\epsilon, u)\}$ are each of order ϵ^2.

This leads to a different interpretation of the advantage offered by the screened estimator. Suppose that, for small ϵ, u, $\Lambda^*(\epsilon) \approx c\epsilon^2$, and, $\max\{\Lambda^*_+(\epsilon, u), \Gamma^*_+(\epsilon, u)\} \approx (c + c')\epsilon^2$, for some $c, c' > 0$. Then for large n, the error of the standard estimator is,

$$\Pr\{S_n > n\epsilon\} \approx e^{-nc\epsilon^2},$$

whereas for the screened estimator,

$$\Pr\{S_n > n\epsilon, \ \ |T_n| < u\} \approx e^{-n(c+c')\epsilon^2}.$$

In both cases, we have approximately Gaussian tails. Therefore, roughly speaking, we may interpret the result of Theorem 5 as saying that, as long as the covariance between $F(X)$ and $U(X)$ is nonzero, *the screened estimates are asymptotically Gaussian with a strictly smaller variance than the standard estimates.*

THEOREM 4. *Suppose that $E[F(X)] = E[U(X)] = 0$ and that $\gamma := Cov(F(X), U(X))$ is nonzero. There exists $\epsilon_0 > 0$ such that, for each $0 < \epsilon < \epsilon_0$, there exists $u_0 = u_0(\epsilon) > 0$ such that $\Delta(\epsilon, u) > 0$ for all $u \in (0, u_0)$.*

Note that the assumption on the covariance being nonzero cannot be relaxed. For example, let $X_i = Y_i Z_i$, $i \geq 1$, where $\{Y_i\}$ are i.i.d. nonnegative random variables, and $\{Z_i\}$ are i.i.d., independent of the $\{Y_i\}$, with each $Z_i = \pm 1$ with probability 1/2. With $F(x) \equiv |x| - E|X_1|$ and $U(X) \equiv \text{sign}(x)$, we have $F(X_i) = Y_i - E(Y_i)$ and $U(X_i) = Z_i$, so that S_n and T_n are independent for all $n \geq 1$. Therefore,

$$\Pr\{S_n > n\epsilon, \ \ |T_n| < nu\} = \Pr\{S_n > n\epsilon\} \Pr\{|T_n| < nu\},$$

and since $\lim_n \Pr\{|T_n| < nu\} = 1$, the exponents of the other two probabilities must be identical.

Whenever γ is nonzero, the variances $\sigma^2(F)$, $\sigma^2(U)$ of $F(X)$ and $U(X)$, respectively, are both nonzero. If $\tilde{\Delta}(\epsilon, u)$ denotes the corresponding difference of exponents for the normalized functions $F/\sigma(F)$ and $U/\sigma(U)$, then from the definitions,

$$\Delta(\epsilon, u) = \tilde{\Delta}\Big(\frac{\epsilon}{\sigma(F)}, \frac{\epsilon}{\sigma(U)}\Big).$$

Therefore, in order to determine the nature of this difference for small ϵ we can assume, without loss of generality, that $\text{Var}(F(X)) = \text{Var}(U(X)) = 1$.

THEOREM 5. *Suppose that $E[F(X)] = E[U(X)] = 0$, that $Var(F(X)) = Var(U(X)) = 1$, and that the covariance $\gamma := Cov(F(X), U(X))$ is nonzero. Then there exists $\alpha > 0$ such that,*

$$\liminf_{\epsilon \to 0} \frac{1}{\epsilon^2}\Delta(\epsilon, \alpha\epsilon) > 0.$$

In fact, there exists $\epsilon_0 > 0$ such that,

$$\Delta\Big(\epsilon, \frac{|\gamma|}{4}\epsilon\Big) \geq \frac{\gamma^2}{8}\epsilon^2,$$

for all $\epsilon \in (0, \epsilon_0)$.

3. SIMULATION

In numerous simulation experiments we found that, when applied to i.i.d. samples, the screening estimator offered no significant advantage over the classical method of control variates [17]. On the other hand, we did observe that in several cases where the underlying samples were produced by a Markov chain, screening was much more effective. Below we present one such example.

Consider the following Bayesian inference problem, as in [17, Example 9.2]. Suppose that we have N independent observations $y = (y_1, y_2, \ldots, y_N)$ from the mixture distribution,

$$pN(\mu_1, \sigma^2) + (1 - p)N(\mu_2, \sigma^2),$$

where the mixing proportion p and the variance σ^2 are assumed to be fixed and known, and that we wish to estimate the value of μ_1. A way to describe this model that facilitates the estimation is to place independent $N(0, 10\sigma^2)$ prior on the means μ_1, μ_2, and introduce latent variables $Z = (Z_1, Z_2, \ldots, Z_N)$, where the Z_i are independent with distribution $P(Z_i = 1) = 1 - P(Z_i = 0) = p$, and, conditional on μ_1, μ_2 and Z, each $Y_i|Z_i = 1 \sim N(\mu_1, \sigma^2)$, and $Y_i|Z_i = 0 \sim N(\mu_2, \sigma^2)$.

The estimation of μ_1 is typically performed by estimating its mean under the *posterior* distribution given the data. In turn, a standard way to do this is via MCMC, as follows. First we note that, under the posterior, conditional on y and z, the parameters μ_1 and μ_2 are independent, with,

$$\begin{aligned}
\pi(\mu_1|y, z) &\sim N\left(\frac{\sum_j z_j y_j}{n_1 + 1/10}, \frac{\sigma^2}{n_1 + 1/10}\right), \\
\pi(\mu_2|y, z) &\sim N\left(\frac{\sum_j (1 - z_j) y_j}{n_2 + 1/10}, \frac{\sigma^2}{n_2 + 1/10}\right),
\end{aligned}$$

respectively, where $n_1 = \sum_j z_j$ is the number of z_i that are equal to 1, and $n_2 = \sum_j (1 - z_j) = N - n_1$ is the number of z_i that are equal to zero. Also, given μ_1, μ_2 and y, the Z_i are independent, and for each $i = 1, 2, \ldots, N$, the posterior probability $\pi(Z_i = 1|\mu_1, \mu_2, y)$ equals,

$$\frac{p \exp\{-(y_i - \mu_1)^2/2\sigma^2\}}{p \exp\{-(y_i - \mu_1)^2/2\sigma^2\} + (1 - p) \exp\{-(y_i - \mu_2)^2/2\sigma^2\}}.$$

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4242
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4242

The *random-scan Gibbs sampler* [17] can be used here to construct a multivariate Markov chain $\{X_n\}$, where each X_i is an $(N+2)$-dimensional vector of the form,

$$(\mu_1, \mu_2, Z) = (\mu_1, \mu_2, Z_1, Z_2, \ldots, Z_N),$$

and where the steady-state distribution of $\{X_n\}$ is exactly the posterior distribution $\pi(\mu_1, \mu_2, Z|y)$. This is done as follows: Start with arbitrary values for $\mu_1(1)$ and $\mu_2(1)$, say $\mu_1(1) = \mu_2(1) = 0$, and draw a sample $Z(1)$ from the conditional distribution described above; this produces the initial value X_1. Then at each step, given X_n, the random-scan Gibbs sampler draws a sample from μ_1, μ_2, or from the entire vector Z, each chosen with probability $1/3$, according to the conditionals of the posterior; X_{n+1} is the same as X_n, except for the one component that has been updated.

Although this Monte Carlo setting is different from the i.i.d. problems we considered so far – since the samples X_i are multivariate, and they are not independent – we can nevertheless define $F(x) = F(\mu_1, \mu_2, z) = \mu_1$, and form the estimates $\{\hat{S}_n\}$ as in the Introduction. The ergodic theorem guarantees that $\hat{S}_n \to \mu_1^*$, the true mean of μ_1 under the posterior, and the associated central limit theorem states that the convergence takes place at a rate $O(n^{-1/2})$.

Models of this type often present a difficultly, in that the posterior on (μ_1, μ_2) is bimodal. As a result, the Gibbs sampler only makes rare transitions between the two modes, and, as a result, the empirical averages $\{\hat{S}_n\}$ have high variability; cf. [17][6]. Figures 5 and 6 show two typical realizations of the Gibbs sampler, illustrating this behavior. The parameter values in these experiments and throughout the remainder of this section, are $N = 300, p = 0.9$ and $\sigma^2 = 1$.

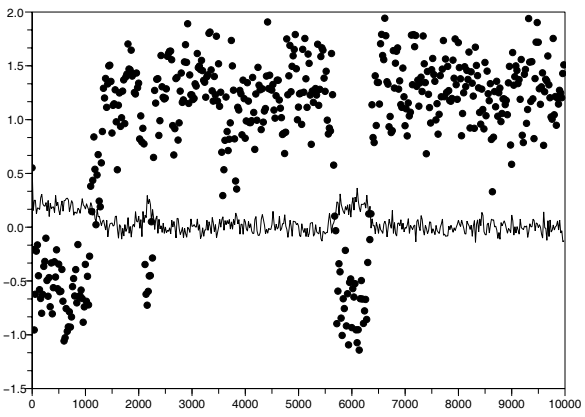

Figure 5: A realization of the samples μ_1, μ_2 produced in $n = 10000$ steps of the Gibbs sampler. The thin solid line is the sequence of μ_1 samples, and the bold points depict the sequence of μ_2 samples. For the sake of visual clarity, the values of μ_2 are plotted only every 20 simulation steps.

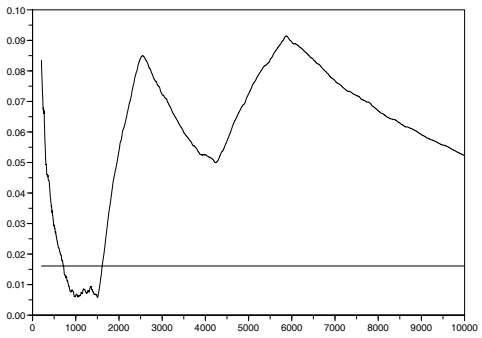

Figure 6: A realization of the sequence of the standard empirical estimates $\{\hat{S}_k\ ;\ 1 \le k \le n\}$ of μ_1, for $n = 10000$ steps of the Gibbs sampler. The horizontal line is the "true" posterior mean of μ_1, estimated to be ≈ 0.0161 after 15 million Gibbs steps.

In order to construct a screening function U, we adopt an idea of Henderson [11]. Let $G(x) = G(\mu_1, \mu_2, z) = \mu_1^2$, and define $U(x) = G(x) - E[X_1|X_0 = x]$, so that,

$$U(x) = \frac{1}{3}\mu_1^2 - \frac{1}{3}\left\{\left(\frac{\sum_j z_j y_j}{n_1 + 1/10}\right)^2 + \frac{\sigma^2}{n_1 + 1/10}\right\}.$$

The definition of U together with the fact that $\pi(\mu_1, \mu_2, Z|y)$ is the stationary distribution of this chain immediately imply that the mean ν of U under π is zero. Therefore, U can be used as a screening function, or, alternatively, we can form the control variate estimates as in Section 1.1, via,

$$\tilde{S}_k := \frac{1}{k}\sum_{i=1}^{k}\Big(F(X_i) - \hat{\beta}U(X_i)\Big) = \hat{S}_k - \hat{\beta}\hat{T}_k,$$

for $1 \le k \le n$, where $\{\hat{T}_k\}$ are the empirical averages of $\{U(X_i)\}$ as in the Introduction, and $\hat{\beta}$ is an adaptive estimate of the optimal coefficient β^*, obtained using a methodology similar to that outlined in [15] and [4]; see these references for details.

Although, because of the high variability of the Gibbs samples it is hard to speak of "typical" instances, we do show one particular realization of all three estimators (the standard empirical averages $\{\hat{S}_k\}$, the control variate estimates $\{\tilde{S}_k\}$, and the screened estimates) in Figure 7.

In order to obtain a more precise idea of the degree of improvement offered by the control variate estimates and by the screening method, for each sample size $n = 100$, 500, 1000, 2000, 5000 and 10000 we performed $T = 500$ repetitions of the same experiment, and we computed the (estimated) variance reduction factor offered by the control variate estimates and by the screened estimates. Specifically, for each repetition $i = 1, 2, \ldots, T = 500$, based on the values of the standard estimates $\hat{S}_n^{(i)}$, we estimated their variance by

$$\sigma^2_{\text{standard}} = \frac{1}{T-1}\sum_{i=1}^{T}[\hat{S}_n^{(i)} - \bar{S}_n]^2,$$

where $\bar{S}_n$ is the average of the values $\hat{S}_n^{(1)}, \hat{S}_n^{(2)}, \ldots, \hat{S}_n^{(T)}$.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4242
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4242

Table 1: Estimated factors by which the variance of the standard empirical averages $\hat{S}_n$ is larger than the corresponding variances of the control-variate estimator and the screened estimator, respectively, after $n = 100, 500, 1000, 2000, 5000$ and 10000 simulation steps.

Variance reduction factors						
	Simulation steps					
Estimator	$n = 100$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$	$n = 10000$
Control variates	< 1	< 1	1.01	<1	1.01	1.00
Screening	1.09	1.23	1.18	1.11	1.16	1.08

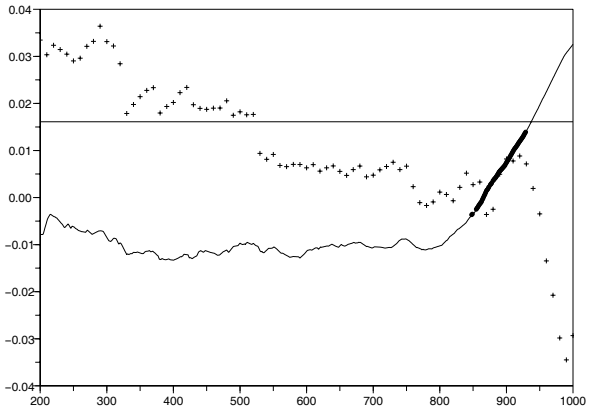

Figure 7: A realization of the three estimators based on sequence $n = 1000$ Gibbs samples. The solid line depicts the standard empirical averages $\{\hat{S}_k\}$ for $200 \leq k \leq 1000$, the "+" signs show the values of the control variate estimates $\{\tilde{S}_k\}$ plotted every 20 steps, and the screened estimates are plotted in bold. The horizontal line is the "true" posterior mean of μ_1, estimated as ≈ 0.0161 after 15 million Gibbs steps.

Similarly, we estimated the variances $\sigma^2_{\text{screening}}$ and σ^2_{cv} of the screening and the control variates estimators, respectively, and the variance reduction factors were estimated by the ratios $\sigma^2_{\text{standard}}/\sigma^2_{\text{cv}}$ and $\sigma^2_{\text{standard}}/\sigma^2_{\text{screening}}$. From the results, shown in Table 1, it is clear that in this scenario screening offers a significant advantage in terms of variance reduction, whereas the control variate estimates fail to produce a meaningful improvement. Finally we note that, although the results in the above example – as well as some other MCMC examples presented in [12] – are quite promising, the actual domain of applicability and the degree of effectiveness of the screened estimator are yet to be precisely determined.

4. ACKNOWLEDGMENTS

We thank Peter Glynn, Jose Blanchet and Petros Dellaportas for several interesting conversations.

5. REFERENCES

[1] S. Asmussen and P. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, New York, 2007.

[2] S. Brooks and A. Gelman. Some issues in monitoring convergence of iterative simulations. In *Proceedings of the Section on Statistical Computing. ASA*, 1998.

[3] I. Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *Ann. Probab.*, 12(3):768–793, 1984.

[4] P. Dellaportas and I. Kontoyiannis. *Manuscript in preparation*, 2008.

[5] A. Dembo and O. Zeitouni. *Large Deviations Techniques And Applications*. Springer-Verlag, New York, second edition, 1998.

[6] J. Diebolt and C. Robert. Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B*, 56(2):363–375, 1994.

[7] G. Fishman. *Monte Carlo: Concepts, Algorithms, and Applications*. Springer-Verlag, New York, 1996.

[8] G. Givens and J. Hoeting. *Computational Statistics*. John Wiley & Sons, Hoboken, NJ, 2005.

[9] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer-Verlag, New York, 2004.

[10] P. Glynn and R. Szechtman. Some new perspectives on the method of control variates. In *Monte Carlo and quasi-Monte Carlo methods, 2000 (Hong Kong)*, pages 27–49. Springer, Berlin, 2002.

[11] S. Henderson. *Variance Reduction Via an Approximating Markov Process*. PhD thesis, Department of Operations Research, Stanford University, Stanford, CA, 1997.

[12] I. Kontoyiannis, L. Lastras-Montaño, and S. Meyn. Explicit exponential deviation bounds for Markov chains via information-theoretic and convex-analytic techniques. *Preprint*, 2008.

[13] I. Kontoyiannis and S. Meyn. Computable exponential bounds for screened estimation and simulation. *To appear, Ann. Appl. Prob.*, 2008. Available online at: pages.cs.aueb.gr/users/yiannisk/.

[14] J. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York, 2001.

[15] S. Meyn. *Control Techniques for Complex Networks*. Cambridge University Press, 2007.

[16] T. Mikosch and A. Nagaev. Large deviations of heavy-tailed sums with applications in insurance. *Extremes*, 1(1):81–110, 1998.

[17] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, second edition, 2004.

[18] I. Sanov. On the probability of large deviations of random variables. *Mat. Sb.*, 42:11-44, 1957. Engl. *Sel. Transl. Math. Statist. Probab.* 1961, **1**, 213-244.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4242
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4242