MATHEMATICS OF OPERATIONS RESEARCH
Vol. 33, No. 1, February 2008, pp. 36–50
ISSN 0364-765X | EISSN 1526-5471 | 08 | 3301 | 0036

DOI 10.1287/moor.1070.0276

Uniformly Efficient Importance Sampling for the Tail Distribution of Sums of Random Variables

Paul Glasserman
Columbia University, New York, New York 10027
pg20@columbia.edu, http://www2.gsb.columbia.edu/faculty/pglasserman/

Sandeep Juneja
Tata Institute of Fundamental Research, Mumbai, India 400005
juneja@tifr.res.in, http://www.tcs.tifr.res.in/~sandeepj/

Successful efficient rare-event simulation typically involves using importance sampling tailored to a specific rare event. However, in applications one may be interested in simultaneous estimation of many probabilities or even an entire distribution. In this paper, we address this issue in a simple but fundamental setting. Specifically, we consider the problem of efficient estimation of the probabilities $P(S_n \geq na)$ for large n, for all a lying in an interval $\mathcal{A}$, where S_n denotes the sum of n independent, identically distributed light-tailed random variables. Importance sampling based on exponential twisting is known to produce *asymptotically efficient* estimates when $\mathcal{A}$ reduces to a single point. We show, however, that this procedure fails to be asymptotically efficient throughout $\mathcal{A}$ when $\mathcal{A}$ contains more than one point. We analyze the best performance that can be achieved using a discrete mixture of exponentially twisted distributions, and then present a method using a continuous mixture. We show that a continuous mixture of exponentially twisted probabilities and a discrete mixture with a sufficiently large number of components produce asymptotically efficient estimates for all $a \in \mathcal{A}$ simultaneously.

Key words: rare-event simulation; importance sampling; large deviations; random walks
MSC2000 subject classification: Primary: 65C05, 60F10, 60G50
OR/MS subject classification: Primary: simulation/efficiency
History: Received February 11, 2006; revised October 31, 2006, December 18, 2006, and December 19, 2006.

1. Introduction. The use of importance sampling for efficient rare-event simulation has been studied extensively (see, e.g., Bucklew [4], Heidelberger [9], Juneja and Shahabuddin [11] for surveys). Application areas for rare-event simulation include communications networks where information loss and large delays are important rare events of interest (as in Chang et al. [6]), insurance risk where the probability of ruin is a critical performance measure (e.g., Asmussen [2]), credit risk models in finance where large losses are a primary concern (e.g., Glasserman and Li [8]), and reliability systems where system failure is a rare event of utmost importance (e.g., Shahabuddin [15]).

Successful applications of importance sampling for rare-event simulation typically focus on the probability of a single rare event. As a way of demonstrating the effectiveness of an importance sampling technique, the probability of interest is often embedded in a sequence of probabilities decreasing to zero. The importance sampling technique is said to be asymptotically efficient or asymptotically optimal if the second moment of the associated estimator decreases at the fastest possible rate as the sequence of probabilities approaches zero.

Importance sampling based on *exponential twisting* produces asymptotically efficient estimates of rare-event probabilities in a wide range of problems. As in the setting we consider here, asymptotic efficiency typically requires that the twisting parameter be tailored to a specific event. However, in applications, one is often interested in estimating many probabilities or an entire distribution. For instance, portfolio credit risk management requires the estimation of the tail of the loss distribution for a range of loss thresholds. This is needed in measuring the amount of capital required to protect against typical and atypical losses, in setting thresholds at which reinsurance treaties are needed to protect against catastrophic losses, and in calculating the premium to be paid for such contracts. Moreover, estimating a tail distribution is often a step toward estimating functions of the tail distribution, such as quantiles or tail conditional expectations. These problems motivate our investigation.

We consider the simple but fundamental setting of tail probabilities associated with a random walk. Exponential twisting produces asymptotically efficient estimates for a single point in the tail; we show, however, that the standard approach fails to produce asymptotically efficient estimates for multiple points simultaneously, a point that has been made in other settings by Bucklew et al. [5], Sadowsky [14], and Siegmund [16]. We develop and analyze modifications that rectify this deficiency. The relatively simple context we consider here provides a convenient setting in which to identify problems and solutions that may apply more generally.

Specifically, we consider the random walk $S_n = \sum_{i=1}^{n} X_i$ where $(X_i\colon i \leq n)$ are independent, identically distributed (iid) random variables with mean μ. We focus on the problem of simultaneous efficient estimation by

simulation of the probabilities $P(S_n/n \geq a)$ for large n, for all $a > \mu$ lying in an interval $\mathcal{A}$. Certain regularity conditions are imposed on this interval. We refer to this problem as that of efficient estimation of the tail distribution curve.

Although the random walk problem we consider is quite simple, the essential features of this problem are often embedded in more complex applications of importance sampling. This holds, for example, in the queueing problem in Sadowsky [14] and the credit risk application in Glasserman and Li [8]. The problem in Glasserman and Li [8] is of the form $P(S_n/n \geq a)$, but in that context the summands producing S_n need not be independent or identically distributed. Nevertheless, the iid case underpins the more general case.

A standard technique for estimating $P(S_n/n \geq a)$, $a > \mu$, uses importance sampling by applying an exponential twist to the (i.i.d.) X_i. It is well known that if the twisting parameter is correctly tailored to a, this method produces asymptotically efficient estimates. (See, e.g., Sadowsky and Bucklew [13].) However, we show in §2.3 that the exponentially twisted distribution that achieves asymptotic efficiency in estimating $P(S_n/n \geq a_1)$ fails to be asymptotically efficient in estimating $P(S_n/n \geq a_2)$ for $a_2 \neq a_1$. In particular, it incurs an exponentially increasing computational penalty as $n \to \infty$. This motivates the need for the alternative procedures that we develop in this paper.

Within the class of exponentially twisted distributions, we first identify the one that estimates the tail distribution curve with asymptotically minimal computational overhead. We then extend this analysis to find the best mixture of $k < \infty$ exponentially twisted distributions. However, even with this distribution, asymptotic efficiency is not achieved.

We note that this shortcoming may be remedied if k is selected to be an increasing function of n. We further propose an importance sampling distribution that is a continuous mixture of exponentially twisted distributions. We show that such a mixture also estimates the tail probability distribution curve asymptotically efficiently for all $a \in \mathcal{A}$ simultaneously. Furthermore, we identify the mixing probability density function that ensures that all points along the curve are estimated with roughly equal precision.

Other settings leading to the simultaneous estimation of multiple performance measures from the same importance sampling distribution include Arsham et al. [1] and Heidelberger et al. [10]. However, the techniques and analysis in these and related work are fundamentally different from those studied in this paper.

The rest of this paper is organized as follows: In §2, we review the basics of importance sampling and introduce some notions of efficiency relevant to our analysis. In §3, we discuss the performance of importance sampling in simultaneously estimating the tail probability distribution curve using discrete mixtures of appropriate exponentially twisted distributions. The analysis is then extended to continuous mixtures in §4. Concluding remarks are given in §5.

2. Naive estimation and importance sampling.

2.1. Naive estimation of the tail distribution curve. Under naive simulation, m iid samples $((X_{i1}, X_{i2}, \ldots, X_{in})\colon i \leq m)$ of $(X_1, X_2, \ldots, X_n)$ are generated using the original probability measure P. Let $S_n^i = \sum_{j=1}^n X_{ij}$. Then, for each $a \in \mathcal{A}$,

$$\frac{1}{m}\sum_{i=1}^{m} I(S_n^i \geq na)$$

provides an unbiased estimator of $P(S_n/n \geq a)$, where $I(\cdot)$ denotes the indicator function of the event in parentheses.

2.2. Importance sampling. We restrict attention to probability measures P^* for which P is absolutely continuous with respect to P^* when both measures are restricted to the σ-algebra generated by $(X_1, X_2, \ldots, X_n)$. Let L denote the likelihood ratio of the restriction of P to the restriction of P^*. Then,

$$P(S_n/n \geq a) = E_{P^*}[LI(S_n/n \geq a)] \tag{1}$$

for each $a \in \mathcal{A}$, where the subscript affixed to E denotes the probability measure used in determining the expectation.

If, under P and P^*, $(X_1, X_2, \ldots, X_n)$ have joint densities f and f^*, respectively, then

$$L = \frac{f(X_1, X_2, \ldots, X_n)}{f^*(X_1, X_2, \ldots, X_n)} \quad \text{a.s.}$$

Clearly, L depends on n, although we do not make this dependence explicit in our notation.

Under importance sampling with probability P^*, m iid samples $((X_{i1}, X_{i2}, \ldots, X_{in})\colon\ i \le m)$ of $(X_1, X_2, \ldots, X_n)$ are generated using P^*. Let $(L_i\colon\ i \le m)$ denote the associated samples of L. For each $a \in \mathcal{A}$, we compute the estimate

$$\frac{1}{m}\sum_{i=1}^{m} L_i I(S_n^i \ge na). \tag{2}$$

In light of (1), (2) provides an unbiased estimator of $P(S_n/n \ge a)$.

The probabilities $P(S_n/n \ge a)$ decrease exponentially in n if the X_i are light tailed. More precisely, if the moment-generating function of the X_i is finite in a neighborhood of the origin, then (as in, e.g., Dembo and Zeitouni [7, p. 34])

$$\lim_{n\to\infty} \frac{1}{n} \log P(S_n/n \ge a) = -\Lambda^*(a), \tag{3}$$

where Λ^* denotes the large deviations rate function associated with the sequence $(S_n/n\colon\ n \ge 0)$, to be defined in §3.2, and a belongs to the interior of $\{x\colon\ \Lambda^*(x) < \infty\}$.

Also note that for any P^*,

$$E_{P^*}[L^2 I(S_n/n \ge a)] \ge (E_{P^*}[LI(S_n/n \ge a)])^2 = P(S_n/n \ge a)^2,$$

and hence

$$\liminf_{n\to\infty} \frac{1}{n} \log E_{P^*}[L^2 I(S_n/n \ge a)] \ge -2\Lambda^*(a). \tag{4}$$

An importance sampling measure P^* is said to be asymptotically efficiently for $P(S_n/n \ge a)$ if equality holds in (4) with the liminf replaced by an ordinary limit.

Furthermore, we say that the *relative variance* of the estimate (under a probability measure P^*) grows polynomially at rate $p > 0$ if

$$0 < \liminf_{n\to\infty} \frac{1}{n^p} \frac{E_{P^*}[L^2 I(S_n/n \ge a)]}{P(S_n/n \ge a)^2} \le \limsup_{n\to\infty} \frac{1}{n^p} \frac{E_{P^*}[L^2 I(S_n/n \ge a)]}{P(S_n/n \ge a)^2} < \infty.$$

It is easily seen that if this holds for some $p > 0$, then P^* is asymptotically efficient for $P(S_n/n \ge a)$. The relative variance of the estimate grows exponentially if the ratio

$$\frac{E_{P^*}[L^2 I(S_n/n \ge a)]}{P(S_n/n \ge a)^2}$$

grows at least at an exponential rate with n.

We say that the relative variance of the estimate of the tail distribution curve over $\mathcal{A}$ grows polynomially at rate $p > 0$ if

$$0 < \sup_{a\in\mathcal{A}} \liminf_{n\to\infty} \frac{1}{n^p} \frac{E_{P^*}[L^2 I(S_n/n \ge a)]}{P(S_n/n \ge a)^2} \le \sup_{a\in\mathcal{A}} \limsup_{n\to\infty} \frac{1}{n^p} \frac{E_{P^*}[L^2 I(S_n/n \ge a)]}{P(S_n/n \ge a)^2} < \infty.$$

It is said to grow at most at the polynomial rate $p > 0$ if the last inequality holds.

2.3. Asymptotically efficient exponentially twisted distribution. We first review the asymptotically efficient exponentially twisted importance sampling distribution for estimating a single probability $P(S_n/n \ge a)$, $a > \mu$. We also show that under this measure the relative variance of the estimate grows polynomially at rate $1/2$.

Some notation and assumptions are needed for this. Let X_i have distribution function F and log-moment generating function Λ under the original probability measure P,

$$\Lambda(\theta) = \log \int \exp(\theta x)\, dF(x).$$

Denote Λ's domain by $\Theta = \{\theta\colon\ \Lambda(\theta) < \infty\}$, and for each $\theta \in \Theta$, let F_θ denote the distribution function obtained by exponentially twisting F by $\theta \in \Theta$, i.e.,

$$dF_\theta(x) = \exp(\theta x - \Lambda(\theta)) dF(x).$$

Let Λ' denote the derivative of Λ. It is a standard consequence of exponential twisting (and easy to see directly) that $\Lambda'(\theta)$ is the mean of a random variable with distribution function F_θ. Set $\mathcal{H} = \{\Lambda'(\theta)\colon\ \theta \in \Theta\}$, i.e., the

collection of possible mean values under distributions obtained by exponentially twisting F. For instance, if F denotes the distribution function of a random variable uniformly distributed between a and b, then $\mathcal{H} = (a, b)$. As another instance, suppose F denotes the distribution function of an exponentially distributed random variable with rate λ. Then, it is easy to check that F_θ corresponds to an exponential distribution with rate $\lambda - \theta$, so $\mathcal{H} = (0, \infty)$.

We assume that Θ has a nonempty interior Θ^o, which then includes zero. Furthermore, we require

$$\mathcal{A} \subseteq \mathcal{H}^o.$$

Let P_θ denote the probability measure under which $(X_i\colon i \geq 1)$ are iid with distribution F_θ. The restrictions of P and P_θ to the σ-algebra generated by $X_1, \ldots, X_n$ are mutually absolutely continuous. Let L_θ denote the Radon-Nikodym derivative of the restriction of P to the restriction of P_θ. Then,

$$L_\theta = \exp(-\theta S_n + n\Lambda(\theta)) \quad \text{a.s.} \tag{5}$$

We recall some definitions associated with the large deviations rate of S_n; see Dembo and Zeitouni [7] for background. The rate function for $(S_n/n\colon n \geq 1)$ is defined by

$$\Lambda^*(a) = \sup_\theta(\theta a - \Lambda(\theta)).$$

For each $a \in \mathcal{H}^o$, there is, by definition of $\mathcal{H}$, a $\theta(a) \in \Theta$ for which $\Lambda'(\theta(a)) = a$. It is easy to see that

$$\theta(a)a - \Lambda(\theta(a)) = \sup_\theta(\theta a - \Lambda(\theta)) = \Lambda^*(a). \tag{6}$$

Moreover, through differentiation, it can be seen that $\Lambda^{*\prime}(a) = \theta(a)$. By differentiating the equation $\Lambda'(\theta(a)) = a$, it follows that

$$\theta'(a) = \frac{1}{\Lambda''(\theta(a))} = \Lambda^{*\prime\prime}(a);$$

as in §2.2.24 of Dembo and Zeitouni [7], we have $\Lambda''(\theta) > 0$ when $\Lambda'(\theta) \in \mathcal{H}^o$. (Furthermore, they note that Λ and Λ^* are infinitely differentiable on Θ^o and $\mathcal{H}^o$, respectively.)

It is well known that (as in Sadowsky and Bucklew [13]) $P_{\theta(a)}$ asymptotically efficiently estimates $P(S_n/n \geq a)$, i.e.,

$$\lim_{n\to\infty} \frac{\log E_{P_{\theta(a)}}[L^2_{\theta(a)} I(S_n/n \geq a)]}{\log P(S_n/n \geq a)} = 2. \tag{7}$$

Theorem 2.1 sharpens this result and also shows what happens when we use a twisting parameter θ different from $\theta(a)$ in estimating $P(S_n/n > a)$. The theorem applies the Bahadur-Rao (Bahadur and Rao [3]) approximations. When X_i has a nonlattice distribution, the Bahadur-Rao approximation states that

$$P(S_n/n \geq a) \sim \frac{1}{\sqrt{2\pi\Lambda''(\theta(a))n}\theta(a)} \exp[-n\Lambda^*(a)]. \tag{8}$$

(We say that $a_n \sim b_n$ for nonnegative sequences $(a_n\colon n \geq 1)$ and $(b_n\colon n \geq 1)$, if $\lim_{n\to\infty} a_n/b_n$ exists and equals 1.) When X_i has a lattice distribution (i.e., for some x_0 and some d, the random variable $d^{-1}(X_i - x_0)$ is a.s. an integer, and d is the largest number with this property) and $0 < P(X_i = a) < 1$, then the Bahadur-Rao approximation states that

$$P(S_n/n \geq a) \sim \frac{1}{\sqrt{2\pi\Lambda''(\theta(a))n}\theta(a)} \exp[-n\Lambda^*(a)] \frac{\theta(a)d}{1 - \exp[-\theta(a)d]}. \tag{9}$$

Let

$$\psi(a, t) = \sqrt{2\pi\Lambda''(\theta(a))}[\theta(a) + \theta(t)].$$

THEOREM 2.1. *For $a > \mu$ and t in $\mathcal{H}^o$:*

(i) *When X_i has a nonlattice distribution:*

$$E_{P_{\theta(t)}}[L^2_{\theta(t)} I(S_n/n \geq a)] \sim \frac{1}{\psi(a, t)\sqrt{n}} \exp[n(\theta(t)(t - a) - \Lambda^*(t) - \Lambda^*(a))]. \tag{10}$$

(ii) *When X_i has a lattice distribution with d as defined above and $P(X_i = a) > 0$:*

$$E_{P_{\theta(t)}}[L^2_{\theta(t)} I(S_n/n \geq a)] \sim \frac{1}{\psi(a,t)\sqrt{n}} \exp[n(\theta(t)(t-a) - \Lambda^*(t) - \Lambda^*(a))] \frac{[\theta(a)+\theta(t)]d}{1-\exp(-[\theta(a)+\theta(t)]d)}. \quad (11)$$

Thus, with $t = a$ the relative variance grows polynomially at rate $1/2$, but with $t \neq a$ the relative variance grows exponentially and $P_{\theta(t)}$ fails to be asymptotically efficient for $P(S_n/n \geq a)$.

PROOF. Once we establish (10) and (11), the polynomial rate of growth of the relative variance when $t = a$ follows from dividing (10) by the square of (8) and dividing (11) by the square of (9). When $t \neq a$, the strict convexity of Λ^* (see Dembo and Zeitouni [7, §2.2.24]) implies that

$$\Lambda^*(a) > \Lambda^*(t) + (a-t)\Lambda^{*\prime}(t).$$

Because $\theta(t) = \Lambda^{*\prime}(t)$, dividing (10) by the square of (8) and dividing (11) by the square of (9) shows that the relative variance grows exponentially in both the cases.

To establish (10) and (11), we use (5) and (6) to write

$$L_{\theta(t)} = \exp\big[-\theta(t)(S_n - na) + n\theta(t)(t-a) - n\Lambda^*(t)\big]. \quad (12)$$

Also note that

$$E_{P_{\theta(t)}}[L^2_{\theta(t)} I(S_n/n \geq a)] = E_P\big[L_{\theta(t)} I(S_n/n \geq a)\big] = E_{P_{\theta(a)}}\big[L_{\theta(t)} L_{\theta(a)} I(S_n/n \geq a)\big].$$

Using (5) to replace $L_{\theta(a)}$ and (12) to replace $L_{\theta(t)}$, we get

$$E_{P_{\theta(t)}}[L^2_{\theta(t)} I(S_n/n \geq a)] = \exp\big[n(\theta(t)(t-a) - \Lambda^*(t) - \Lambda^*(a))\big] E_{P_{\theta(a)}}\big[\exp[-(\theta(t)+\theta(a))(S_n - na)] I(S_n/n \geq a)\big].$$

Using exactly the same arguments as in Bahadur and Rao [3] or Dembo and Zeitouni [7, §3.7.4], it follows that

$$E_{P_{\theta(a)}}\big[\exp[-(\theta(t)+\theta(a))(S_n - na)] I(S_n/n \geq a)\big] \sim \frac{1}{\psi(a,t)\sqrt{n}},$$

when X_i has a nonlattice distribution, and

$$E_{P_{\theta(a)}}\big[\exp[-(\theta(t)+\theta(a))(S_n - na)] I(S_n/n \geq a)\big] \sim \frac{1}{\psi(a,t)\sqrt{n}} \frac{[\theta(a)+\theta(t)]d}{1-\exp(-[\theta(a)+\theta(t)]d)},$$

when X_i has a lattice distribution. □

Other results showing the uniqueness of an asymptotically efficient exponential twist are established in Bucklew et al. [5], Sadowsky [14], and Siegmund [16].

3. Finite mixtures of exponentially twisted distributions. We have seen in the previous section that a single exponential change of measure cannot estimate multiple points along the tail distribution curve with asymptotic efficiency. In this section, we identify the twisting parameter that is minimax optimal in the sense that it minimizes the worst-case relative variance over an interval $\mathcal{A} = [a_1, a_2] \subseteq \mathcal{H}^o$, $a_1 > \mu$. We then consider an optimal mixture of $k < \infty$ exponentially twisted distributions. We show that if $k \to \infty$ as $n \to \infty$, then the tail distribution curve is asymptotically efficiently estimated over $\mathcal{A}$. Furthermore, if k is $\underline{\Theta}(\sqrt{n/\log n})$[1] the relative variance of the tail probability distribution curve grows polynomially over $\mathcal{A}$.

Some definitions are useful for our analysis. From Theorem 2.1, it follows that for $a \in \mathcal{H}^o$ and $a > \mu$,

$$\lim_{n\to\infty} \frac{1}{n} \log E_{P_{\theta(t)}}[L^2_{\theta(t)} I(S_n/n \geq a)] = -H(t,a),$$

where (recalling that $\theta(t) = \Lambda^{*\prime}(t)$),

$$H(t,a) = \Lambda^*(a) + \Lambda^*(t) + \Lambda^{*\prime}(t)(a-t).$$

[1] A nonnegative function $f(x)$ is said to be $\underline{\Theta}(g(x))$, where g is another nonnegative function, if there exists a positive constant K such that $f(x) \geq Kg(x)$ for all x sufficiently large.

The asymptotic rate of the relative variance of the estimator under $P_{\theta(t)}$ may be defined as

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\frac{E_{P_{\theta(t)}}[L^2_{\theta(t)}I(S_n/n\ge a)]}{P(S_n/n\ge a)^2}\right).$$

From Theorem 2.1, this equals

$$J(t,a)=\Lambda^*(a)-(a-t)\Lambda^{*\prime}(t)-\Lambda^*(t)=2\Lambda^*(a)-H(t,a).$$

This may be interpreted as the nonnegative exponential penalty incurred for twisting with parameter $\theta(t)$ in estimating $P(S_n/n\ge a)$. It equals zero at $t=a$ and is positive otherwise. The penalty $J(t,a)$ has the following geometric interpretation: It equals the vertical distance between the rate function Λ^* and the tangent drawn to this function at t, both evaluated at a.

3.1. The minimax optimal twist. A formulation of the problem of minimizing the worst-case asymptotic relative variance is to find $t\in\mathcal{H}^o$ minimizing

$$\sup_{a\in\mathcal{A}}\lim_{n\to\infty}\frac{1}{n}\log\left(\frac{E_{P_{\theta(t)}}[L^2_{\theta(t)}I(S_n/n\ge a)]}{P(S_n/n\ge a)^2}\right)=\sup_{a\in\mathcal{A}}J(t,a). \tag{13}$$

Proposition 3.1. *For $\mathcal{A}=[a_1,a_2]\subseteq\mathcal{H}^o$, $a_1>\mu$, the asymptotic worst-case relative variance* (13) *is minimized over t by t^* that satisfies*:

$$\Lambda^{*\prime}(t^*)=\frac{\Lambda^*(a_2)-\Lambda^*(a_1)}{a_2-a_1}. \tag{14}$$

Proof. The optimization problem reduces to

$$\inf_{t\in\mathcal{H}^o}\sup_{a\in[a_1,a_2]}(\Lambda^*(a)-\Lambda^{*\prime}(t)(a-t)-\Lambda^*(t)).$$

Due to the convexity of Λ^*, for any t, $\sup_{a\in[a_1,a_2]}(\Lambda^*(a)-\Lambda^{*\prime}(t)(a-t)-\Lambda^*(t))$ is achieved at a_1 or a_2. Thus, the problem reduces to

$$\inf_{t\in\mathcal{H}^o}\max\{\Lambda^*(a_1)-\Lambda^{*\prime}(t)(a_1-t)-\Lambda^*(t),\Lambda^*(a_2)-\Lambda^{*\prime}(t)(a_2-t)-\Lambda^*(t)\}.$$

If $t<a_1$, then both functions inside the max are decreasing in t, so increasing t reduces the maximum. If $t>a_2$, then both functions inside the max are increasing in t, so decreasing t reduces the maximum. It therefore suffices to consider $t\in[a_1,a_2]$. At all such t, the first function is increasing and the second is decreasing, so the maximum is minimized where they are equal, which is the point t^* in (14). □

3.2. Finite mixtures: Minimax objective. We now consider a probability measure $\tilde{P}$ that is a nonnegative mixture of k exponentially twisted distributions so that

$$\tilde{P}(A)=\sum_{i=1}^k p_iP_{\theta(t_i)}(A),$$

where $p_i>0$, $\sum_{i=1}^k p_i=1$, and $t_1<t_2<\cdots<t_k$, and each $t_i\in\mathcal{H}^o$. Our objective is to find the optimal twisting parameters for a fixed k. Although our analysis remains valid for all $p_i>0$, $\sum_{i=1}^k p_i=1$, we later note that from a practical viewpoint, $p_i=1/k$ for each i is a reasonable choice when $\tilde{P}$ uses optimal twisting parameters. We then examine the asymptotic relative variance as k increases. Mixtures of exponentially twisted distributions have been used in many applications, including Sadowsky and Bucklew [13].

For a fixed k, we formulate the problem of selecting a mixture of k exponentially twisted distributions to minimize the worst-case asymptotic relative variance over $\mathcal{A}=[a_1,a_2]\subseteq\mathcal{H}^o$ as follows:

$$\inf_{t_1,\ldots,t_k\in\mathcal{H}^o}\sup_{a\in\mathcal{A}}\lim_{n\to\infty}\frac{1}{n}\log\left(\frac{E_{\tilde{P}}[\tilde{L}^2I(S_n/n\ge a)]}{P(S_n/n\ge a)^2}\right), \tag{15}$$

where $\tilde{L}$ is the likelihood ratio of P with respect to $\tilde{P}$, given by

$$\tilde{L}=\frac{1}{\sum_{i=1}^k p_i\exp[\theta(t_i)S_n-n\Lambda(\theta(t_i))]}.$$

The existence of the limit in (15) is guaranteed by the following lemma. Recall that $H(t,a)=\Lambda^*(a)+\Lambda^*(t)+\Lambda^{*\prime}(t)(a-t)$.

LEMMA 3.1. *For $a, t_1, \ldots, t_k \in \mathcal{H}^o$, $a > \mu$, there exists a constant c such that*

$$E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)] \leq c \exp\Big(-n \max_{i=1,\ldots,k} H(t_i, a)\Big) \tag{16}$$

for all sufficiently large n. Furthermore,

$$\lim_{n\to\infty} \frac{1}{n} \log E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)] = -\max_{i=1,\ldots,k} H(t_i, a). \tag{17}$$

Recall that $J(t, a) = 2\Lambda^*(a) - H(t, a)$. In view of (17), our optimization problem reduces to

$$\inf_{t_1,\ldots,t_k \in \mathcal{H}^o} \sup_{a \in \mathbb{R}} \min_{i=1,\ldots,k} J(t_i, a). \tag{18}$$

PROOF OF LEMMA 3.1. We first show (16). We have

$$E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)] = E_P[\tilde{L} I(S_n/n \geq a)] \tag{19}$$

and

$$\tilde{L} \leq \min_{i=1,\ldots,k} (1/p_i) \exp[-\theta(t_i) S_n + n\Lambda(\theta(t_i))],$$

so

$$E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)] \leq \min_{i=1,\ldots,k} (1/p_i) \exp[-\theta(t_i) na + n\Lambda(\theta(t_i))] P(S_n/n \geq a).$$

From (8) we get

$$P(S_n/n \geq a) \leq \text{constant} \cdot \exp[-n\Lambda^*(a)]$$

for all sufficiently large n. Recalling that $\theta(t_i) = \Lambda^{*\prime}(t_i)$, and $\Lambda^*(t) = \theta(t)t - \Lambda(\theta(t))$, we therefore get

$$E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)] \leq \text{constant} \cdot \min_{i=1,\ldots,k} (1/p_i) \exp[-nH(t_i, a)] \tag{20}$$

for all sufficiently large n. This proves the upper bound in the lemma because the minimum is achieved by the largest exponent $H(t_i, a)$, $i = 1, \ldots, k$, for all sufficiently large n.

To see (17), from (20) it follows that

$$\limsup_{n\to\infty} \frac{1}{n} \log E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)] \leq \min_{i=1,\ldots,k} -H(t_i, a) = -\max_{i=1,\ldots,k} H(t_i, a).$$

To get a lower bound, choose $\epsilon > 0$ and write

$$E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)] = E_{P_{\theta(a+\epsilon)}}[\tilde{L} L_{\theta(a+\epsilon)} I(S_n/n \geq a)] \geq E_{P_{\theta(a+\epsilon)}}[\tilde{L} L_{\theta(a+\epsilon)} I(a + 2\epsilon \geq S_n/n \geq a)].$$

This last expression is, in turn, bounded below by

$$\begin{aligned} \widetilde{M}_{n,\epsilon} &= \exp[-\theta(a+\epsilon)(a+2\epsilon)n + n\Lambda(\theta(a+\epsilon))] \\ &\quad \times \left(\sum_{i=1}^{k} p_i \exp(\theta_i(a+2\epsilon)n - n\Lambda(\theta_i))\right)^{-1} P_{\theta(a+\epsilon)}(a \leq S_n/n \leq a+2\epsilon), \end{aligned}$$

where we have written θ_i for $\theta(t_i)$. Because the X_i have mean $a+\epsilon$ under $P_{\theta(a+\epsilon)}$,

$$P_{\theta(a+\epsilon)}(a \leq S_n/n \leq a+2\epsilon) \to 1.$$

Thus,

$$\liminf_{n\to\infty} \frac{1}{n} \log \widetilde{M}_{n,\epsilon} \geq -\theta(a+\epsilon)(a+2\epsilon) + \Lambda(\theta(a+\epsilon)) - \max_{i=1,\ldots,k} \{\theta_i(a+2\epsilon) - \Lambda(\theta_i)\}$$

from which follows

$$\liminf_{n\to\infty} \frac{1}{n} \log E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)] \geq -\theta(a+\epsilon)(a+2\epsilon) + \Lambda(\theta(a+\epsilon)) - \max_{i=1,\ldots,k} \{\theta_i(a+2\epsilon) - \Lambda(\theta_i)\}.$$

Because $\epsilon > 0$ is arbitrary, we also have

$$\liminf_{n\to\infty} \frac{1}{n} \log E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)] \geq -\theta(a)a + \Lambda(\theta(a)) - \max_{i=1,\ldots,k} \{\theta_i a - \Lambda(\theta_i)\} = -\max_{i=1,\ldots,k} H(t_i, a). \quad \square$$

3.3. Finite mixtures: Optimal parameters. We now turn to the solution of (18), the problem of finding the optimal twisting parameters. We first develop some simple results related to convex functions that are useful for this.

Consider a differentiable function f that is strictly convex on an interval that includes $[y, z]$ in its interior. Consider the problem of finding the maximum vertical distance between the graph of f and the line joining the points $(y, f(y))$ and $(z, f(z))$; i.e.,

$$g(y,z) = \max_{t\in[y,z]} \left(f(z) - (z-t)\frac{f(z)-f(y)}{z-y} - f(t) \right).$$

At the optimal point t^*, we have

$$f'(t^*) = \frac{f(z)-f(y)}{z-y}, \tag{21}$$

and

$$g(y,z) = f(z) - (z-t^*)f'(t^*) - f(t^*). \tag{22}$$

By substituting for $f'(t^*)$, it is also easily seen that

$$g(y,z) = f(y) + (t^*-y)f'(t^*) - f(t^*). \tag{23}$$

Also note that $g(y, z)$ is a continuous function of (y, z); it increases in z and decreases in y.

The following lemma is useful in arriving at optimal parameters of finite mixtures. Its content is intuitively obvious: Given a strictly convex function on an interval, the interval can be partitioned into k subintervals such that the maximum distance between the function and the piecewise-linear interpolation between the endpoints of the subintervals is the same over all subintervals.

LEMMA 3.2. *Suppose that f is strictly convex on $[y, z]$. Then there exist points $y = b_1 < b_2 < \cdots < b_{k+1} = z$ satisfying*

$$g(b_i, b_{i+1}) = g(b_{i+1}, b_{i+2}) \quad i = 1, 2, \ldots, k-1. \tag{24}$$

Note that whenever the existence of $(b_1, b_2, \ldots, b_{k+1})$ as in Lemma 3.2 is established, the common value of $g(b_i, b_{i+1})$ for $i = 1, 2, \ldots, k$, is determined by b_{k+1} if b_1 is held fixed. In the proof, $J_k(b_{k+1})$ denotes this value.

PROOF OF LEMMA 3.2. The proof is by induction. First consider $k = 2$. Here, $g(b_1, b) - g(b, b_3)$ is a continuous increasing function of b. It is negative at $b = b_1$ and positive at $b = b_3$. Thus, there exists $b_2 \in (b_1, b_3)$ such that $g(b_1, b_2) = g(b_2, b_3) = J_2(b_3)$. Furthermore, as b_3 increases, $g(b_1, b) - g(b, b_3)$ decreases, so b_2 and hence $J_2(b_3)$ increase continuously with b_3.

To proceed by induction, assume that for each $b_k > y$, there exist points $y = b_1 < b_2 < \cdots < b_k$ such that $g(b_i, b_{i+1}) = g(b_{i+1}, b_{i+2}) = J_{k-1}(b_k)$ for $i = 1, 2, \ldots, k-2$. Furthermore, each b_i and $J_{k-1}(b_k)$ is an increasing function of b_k. Now consider the function $J_{k-1}(b) - g(b, b_{k+1})$ as a function of b. This function is negative at $b = b_1$, positive at $b = b_{k+1}$, and it increases continuously with b. So, again there exists a b_k so that $J_{k-1}(b_k) = g(b_k, b_{k+1})$. Set this value equal to $J_k(b_{k+1})$. Again, it may similarly be seen that b_k, and hence all b_i, $i = 2, 3, \ldots, k-1$, and $J_k(b_{k+1})$ increase with b_{k+1}. □

Given points $(b_1, b_2, \ldots, b_{k+1})$, as in Lemma 3.2, we may define t_i by

$$f'(t_i) = \frac{f(b_{i+1}) - f(b_i)}{b_{i+1} - b_i}, \quad i = 1, 2, \ldots, k. \tag{25}$$

Then, from (22) and (23), it is easy to see that (24) amounts to

$$f(t_{i+1}) - f(t_i) = f'(t_i)[b_{i+1} - t_i] + f'(t_{i+1})[t_{i+1} - b_{i+1}], \quad i = 1, 2, \ldots, k-1.$$

In Proposition 3.2, we use these observations and Lemma 3.2 to identify the solution of (18), i.e., the optimal twisting parameters.

PROPOSITION 3.2. *Suppose that $\mathcal{A} = [a_1, a_2] \subseteq \mathcal{H}^o$, $a_1 > \mu$. Let the points $a_1 = b_1 < b_2 < \cdots < b_{m+1} = a_2$ and $t_1, \ldots, t_k$ satisfy*

$$\Lambda^{*\prime}(t_i) = \frac{\Lambda^*(b_{i+1}) - \Lambda^*(b_i)}{b_{i+1} - b_i}, \quad i = 1, \ldots, k \tag{26}$$

and

$$\Lambda^*(t_{i+1}) - \Lambda^*(t_i) = \Lambda^{*\prime}(t_i)[b_{i+1} - t_i] + \Lambda^{*\prime}(t_{i+1})[t_{i+1} - b_{i+1}], \quad i = 1, \ldots, k-1, \tag{27}$$

as in Lemma 3.2 and (24). These $t_1, \ldots, t_k$ solve (18).

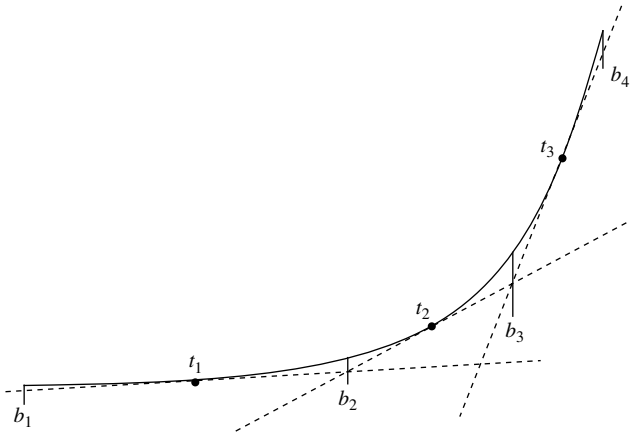

FIGURE 1. Illustration of the points t_i and b_i.

PROOF. From the strict convexity of Λ^* in $\mathcal{H}^o$, it follows that the tangent lines at $t_1, \ldots, t_k$ partition the interval $[a_1, a_2]$ into k subintervals such that throughout the ith interval, Λ^* is closer to the tangent line at t_i than it is to the tangent line at any other t_j. The endpoints of the ith subinterval, $i = 2, \ldots, k-1$, are the points at which the tangent line at t_i crosses the tangent lines at t_{i-1} and t_{i+1}. These are the points $b_2, \ldots, b_m$ in (27); see Figure 1. The left endpoint of the first subinterval is simply $a_1 = b_1$ and the right endpoint of the mth subinterval is simply $a_2 = b_{k+1}$. Thus, for all $a \in [b_i, b_{i+1}]$,

$$\min_{j=1,\ldots,k} J(t_j, a) = J(t_i, a).$$

Furthermore, each $J(t_i, \cdot)$ is a convex function, so

$$\sup_{b_i \le a \le b_{i+1}} J(t_i, a) = \max\{J(t_i, b_i), J(t_i, b_{i+1})\}, \quad i = 1, \ldots, k.$$

Viewing $b_2, \ldots, b_k$ as functions of $t_1, \ldots, t_k$ (through (27)), the problem in (18) thus becomes one of minimizing

$$\max\{J(t_1, b_1), J(t_1, b_2), J(t_2, b_2), J(t_2, b_3), \ldots, J(t_k, b_k), J(t_k, b_{k+1})\}$$

over $t_1, \ldots, t_k$. In fact, (27) says that $J(t_i, b_{i+1}) = J(t_{i+1}, b_{i+1})$, $i = 1, \ldots, k-1$, so we may simplify this to minimizing

$$\max\{J(t_1, b_1), J(t_2, b_2), \ldots, J(t_k, b_k), J(t_k, b_{k+1})\}. \tag{28}$$

Each b_i, $i = 2, \ldots, k$, is an increasing function of t_{i-1} and t_i. Each $J(t_i, b_i) \equiv J(t_i, b_i(t_{i-1}, t_i))$ is easily seen to be an increasing function of t_i and a decreasing function of t_{i-1}, $i = 2, \ldots, k$. Similarly, $J(t_1, b_1)$ is an increasing function of t_1 and $J(t_k, b_{k+1})$ is a decreasing function of t_k. It follows that if all values inside the max in (28) are equal, then $t_1, \ldots, t_k$ are optimal: No change in $t_1, \ldots, t_k$ can simultaneously reduce all values inside the max.

All values in (28) are equal if $J(t_i, b_i) = J(t_{i+1}, b_{i+1})$, $i = 1, \ldots, k-1$, and $J(t_k, b_k) = J(t_k, b_{k+1})$. Simple algebra now shows that this is equivalent to (26). □

The proof of Proposition 3.2 leads to a simple algorithm for finding the optimal points $t_1, \ldots, t_k$ to use in a mixture of m exponentially twisted distributions. The key observation is that it suffices to find the optimal value of (28), which (as noted in the proof) is attained when all values inside the maximum in (28) are equal. Start with a guess $\epsilon > 0$ for this value and set $b_1 = a_1$. Now proceed recursively, for $i = 1, \ldots, m$: Given b_i, find t_i by solving the equation

$$\Lambda^*(b_i) - \Lambda^*(t_i) - \Lambda^{*\prime}(t_i)[b_i - t_i] = \epsilon$$

and then find b_{i+1} by solving

$$\Lambda^*(b_{i+1}) - \Lambda^*(t_i) - \Lambda^{*\prime}(t_i)[b_{i+1} - t_i] = \epsilon.$$

This is straightforward, because the left side of the first equation is increasing in the unknown t_i and the left side of the second equation is increasing in the unknown b_{i+1}. If at some step there is no solution in $[a_1, a_2]$, the procedure stops, reduces ϵ, and starts over at b_1. If a solution is found at every step but $b_{k+1} < a_2$, then ϵ is increased and the procedure starts over at b_1. Thus, one may apply a bisection search over ϵ to find the optimal ϵ and, simultaneously, the optimal t_i and b_i. The exact optimum has $b_{k+1} = a_2$; in practice, one must specify some error tolerance for $|b_{k+1} - a_2|$.

REMARK 3.1 (SELECTING p_iS). Let J denote the minimal value in (18), the minimax penalty incurred in estimating $P(S_n/n \geq a)$ for all $a \in \mathcal{A}$ using a mixture of k exponentially twisted distributions. Suppose that a mixture distribution $\tilde{P}$ with optimal twisting parameters and mixing weights $(p_1, \ldots, p_k)$ is used. We now argue that $p_i = 1/k$ for all i is a reasonable choice of weights.

Note that because $\mathcal{A}$ is compact, there exists a constant $\tilde{c} > 0$ so that for all $a \in \mathcal{A}$,

$$P(S_n/n \geq a) \geq \frac{\tilde{c}}{\sqrt{n}} \exp[-n\Lambda^*(a)] \tag{29}$$

for all sufficiently large n.

From (20) and (29) it follows that

$$\frac{E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)]}{P(S_n/n \geq a)^2} \leq \frac{n}{\tilde{c}^2} \cdot \min_{j=1,\ldots,k} (1/p_j) \exp[nJ(t_j, a)] \tag{30}$$

for all sufficiently large n. For $a \in [b_i, b_{i+1}]$, we have $J(t_i, a) = \min_{j=1,\ldots,k} J(t_j, a) \leq J$. Therefore, for n sufficiently large,

$$\frac{E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)]}{P(S_n/n \geq a)^2} \leq \frac{n}{\tilde{c}^2} \cdot (1/p_i) \exp[nJ(t_i, a)] \leq \frac{n}{\tilde{c}^2} \cdot (1/p_i) \exp[nJ],$$

so that

$$\sup_{a \in \mathcal{A}} \frac{E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)]}{P(S_n/n \geq a)^2} \leq \frac{n}{\tilde{c}^2} \exp[nJ] \max_{i=1,\ldots,k} (1/p_i).$$

Hence, a reasonable choice for the p_is minimizes this upper bound, i.e., solves the minimax problem $\min \max_{i=1,\ldots,k} (1/p_i)$ subject to $\sum_{i \leq k} p_i = 1$. This corresponds to $p_i = 1/k$ for each i. A more refined choice of p_is may be made by using the exact asymptotic (8) instead of the lower bound (29) in the above discussion.

3.4. Increasing the number of components. Again consider $\mathcal{A} = [a_1, a_2] \subseteq \mathcal{H}^o$ and let J_k denote the minimal value in (18), where we now explicitly show its dependence on k. We examine how J_k decreases with k. Recall that for $a \in \mathcal{H}^o$, $0 < \Lambda''(\theta(a)) < \infty$. Because $\Lambda^{*\prime\prime}(a) = 1/\Lambda''(\theta(a))$, it follows that $0 < \Lambda^{*\prime\prime}(a) < \infty$. Because $\mathcal{A} \subseteq \mathcal{H}^o$ is a compact set, we have $\inf_{a \in \mathcal{A}} \Lambda^{*\prime\prime}(a) > 0$ and $\sup_{a \in \mathcal{A}} \Lambda^{*\prime\prime}(a) < \infty$.

PROPOSITION 3.3. *For $\mathcal{A} = [a_1, a_2] \subseteq \mathcal{H}^o$, $a_1 > \mu$, there exist constants d_1, d_2 such that for all $k = 1, 2, \ldots,$*

$$\frac{d_1}{k^2} \leq J_k \leq \frac{d_2}{k^2}.$$

PROOF. Let $c_- = \inf_{a \in \mathcal{A}} \Lambda^{*\prime\prime}(a)$ and $c_+ = \sup_{a \in \mathcal{A}} \Lambda^{*\prime\prime}(a)$. For any interval $[b_i, b_{i+1}] \subseteq \mathcal{A}$ and point t_i satisfying (26), twice integrating the bounds on the second derivative of $\Lambda^{*\prime\prime}$ yields

$$\frac{c_-}{2}(t_i - b_i)^2 \leq \Lambda^*(b_i) - \Lambda^*(t_i) - \Lambda^{*\prime}(t_i)[b_i - t_i] \leq \frac{c_+}{2}(t_i - b_i)^2$$

and

$$\frac{c_-}{2}(t_i - b_{i+1})^2 \leq \Lambda^*(b_{i+1}) - \Lambda^*(t_i) - \Lambda^{*\prime}(t_i)[b_{i+1} - t_i] \leq \frac{c_+}{2}(t_i - b_{i+1})^2.$$

To construct an upper bound on J_k, we may take $b_1, \ldots, b_{k+1}$ to be equally spaced, so that $b_{i+1} - b_i = (a_2 - a_1)/k$, and choose t_i to satisfy (26). Then

$$J_k \leq \max_{i=1,\ldots,k} \sup_{b_i \leq a \leq b_{i+1}} \{\Lambda^*(a) - \Lambda^*(t_i) - \Lambda^{*\prime}(t_i)[a - t_i]\}.$$

As in the proof of Proposition 3.1, the supremum over the ith subinterval is attained at the endpoints, so

$$J_k \leq \max_{i=1,\ldots,k} \frac{c_+}{2}(t_i - b_i)^2 \leq \frac{c_+(a_2 - a_1)^2}{2k^2}.$$

To get a lower bound on J_k, observe that for any $b_1 < \cdots < b_{k+1}$ with $a_1 = b_1$ and $b_{k+1} = a_2$ (including the optimal values), at least one subinterval $[b_i, b_{i+1}]$ must have length greater than or equal to $(a_2 - a_1)/k$. Fix such a subinterval and let t_i satisfy (26). Then,

$$J_k \geq \max\{\Lambda^*(b_i) - \Lambda^*(t_i) - \Lambda^{*\prime}(t_i)[b_i - t_i], \Lambda^*(b_{i+1}) - \Lambda^*(t_i) - \Lambda^{*\prime}(t_i)[b_{i+1} - t_i]\},$$

so

$$J_k \geq \max\left\{\frac{c_-}{2}(t_i - b_{i+1})^2, \frac{c_-}{2}(t_i - b_i)^2\right\} \geq \frac{c_-}{8}(b_{i+1} - b_i)^2 \geq \frac{c_-}{8k^2}(a_2 - a_1)^2. \quad \square$$

The following result shows that by increasing the number of components k in the mixture together with n, we can recover asymptotic efficiency. Here $\tilde{P}$ is defined from k components, as before, using the optimal twisting parameters $\theta(t_i)$, $i = 1, \ldots, k$.

PROPOSITION 3.4. *For $\mathcal{A} = [a_1, a_2] \subseteq \mathcal{H}^o$, $a_1 > \mu$, if $k \to \infty$ as $n \to \infty$, then $\tilde{P}$ is asymptotically efficient for $P(S_n/n \geq a)$, for every $a \in \mathcal{A}$. If $k = \underline{\Theta}(\sqrt{n/\log n})$, then the relative variance of the estimated tail distribution curve over $\mathcal{A}$ is polynomially bounded,*

$$\sup_{a \in \mathcal{A}} \limsup_{n \to \infty} \frac{1}{n^p} \frac{E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)]}{P(S_n/n \geq a)^2} < \infty, \tag{31}$$

for $p > 1$. If $k = \underline{\Theta}(\sqrt{n})$, then (31) holds with $p = 1$.

PROOF. From Lemma 3.1 and (8), we know that

$$E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)] \leq \text{constant} \cdot \exp\Big(-n \max_{i=1,\ldots,k} H(t_i, a)\Big)$$

and

$$P(S_n/n \geq a) \geq \frac{\text{constant}}{\sqrt{n}} \cdot \exp(-n\Lambda^*(a))$$

for all sufficiently large n. Thus,

$$\begin{aligned} \frac{E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)]}{P(S_n/n \geq a)^2} &\leq n \cdot \text{constant} \cdot \exp\Big(n\Big[2\Lambda^*(a) - \max_{i=1,\ldots,k} H(t_i, a)\Big]\Big) \\ &\leq n \cdot \text{constant} \cdot \exp(nJ_k) \leq n \cdot \text{constant} \cdot \exp(nd_2/k^2) \end{aligned} \tag{32}$$

for all sufficiently large n. If $k \to \infty$, then

$$\lim_{n \to \infty} \frac{1}{n} \log\left(\frac{E_{\tilde{P}}[\tilde{L}^2 I(S_n/n \geq a)]}{P(S_n/n \geq a)^2}\right) = 0.$$

From this, asymptotic efficiency as an estimator of $P(S_n/n \geq a)$ follows.

If $k = \underline{\Theta}(\sqrt{n/\log n})$, (32) implies (31) for any $p > 1$. If $k = \underline{\Theta}(\sqrt{n})$, (32) implies (31) with $p = 1$. □

Bounds of the same order of magnitude as those in Proposition 3.3 (but with different constants) hold if we use equally spaced points t_i rather than the optimal points. Thus, the number of components k needed to achieve asymptotic efficiency in that case is of the same order of magnitude as with optimally chosen points. This suggests that choosing the t_i optimally is most relevant when k is relatively small.

REMARK 3.2 ($\mathcal{A}$ IS FINITE). Our primary focus in the paper has been simultaneously efficiently estimating the probabilities $P(S_n/n \geq a)$ for each a lying in an interval $\mathcal{A}$. It is worth noting that the problem is quite simple when $\mathcal{A}$ is a finite set. To see this, suppose that $\mathcal{A} = \{a_1, a_2, \ldots, a_k\}$, where each $a_i > \mu$. To simultaneously efficiently estimate probabilities $(P(S_n/n \geq a_i)\colon i \leq k)$, it is natural to consider the probability measure

$$\hat{P}(A) = \sum_{i=1}^{k} p_i P_{\theta(a_i)}(A).$$

To see that this asymptotically optimally estimates $P(S_n/n \geq a_i)$, note that the associated likelihood ratio

$$\hat{L} = \frac{1}{\sum_{i=1}^{k} p_i \exp[\theta(a_i)S_n - n\Lambda(\theta(a_i))]}$$

on the set $\{S_n/n \geq a_i\}$ is upper bounded by

$$\frac{1}{p_i} \exp(-n\Lambda^*(a_i)).$$

Hence,

$$E_{\hat{P}}[\hat{L}^2 I(S_n/n \geq a_i)] \leq \frac{1}{p_i^2} \exp(-2n\Lambda^*(a_i))$$

and asymptotic optimality of $\hat{P}$ follows.

4. Continuous mixture of exponentially twisted distributions. We now show that a continuous mixture of $P_{\theta(t)}$ for $t \in \mathcal{A}$, simultaneously asymptotically efficiently estimates each $P(S_n/n \geq a)$ for each $a \in \mathcal{A}$. In Theorem 4.1, we allow $\mathcal{A}$ to be any interval in $\mathcal{H}^o$ of values greater than μ. Thus, it is allowed to have the form $[a_1, \infty)$ as long as it is a subset of $\mathcal{H}^o$ and $a_1 > \mu$.

4.1. Polynomially bounded relative variance. Let g be a probability density function with support $\mathcal{A}$. Consider the probability measure P_g, where for any set A,

$$P_g(A) = \int_{a_1}^{a_2} P_{\theta(t)}(A) g(t)\, dt.$$

(Here a_2 may equal ∞.) This measure may be used to estimate $P(S_n/n \geq a)$ as follows: First, a sample T is generated using the density g. Then, the exponentially twisted distribution $P_{\theta(T)}$ is used to generate the sample $(X_1, X_2, \ldots, X_n)$, and hence the output is $L_g(S_n) I(S_n/n \geq a)$ where

$$L_g(S_n) = \left(\int_{a_1}^{a_2} \exp[\theta(t) S_n - n\Lambda(\theta(t))] g(t)\, dt \right)^{-1}. \tag{33}$$

THEOREM 4.1. *For each $a \in \mathcal{A}^o$, when g is a positive continuous function on $\mathcal{A}$,*

$$\limsup_{n\to\infty} \frac{1}{n} \frac{E_{P_g}[L_g(S_n)^2 I(S_n/n \geq a)]}{P(S_n/n \geq a)^2} \leq \frac{\theta(a)\Lambda''(\theta(a))}{g(a)} = \frac{\Lambda^{*\prime}(a)}{\Lambda^{*\prime\prime}(a) g(a)},$$

so that the relative variance of the associated importance sampling estimator grows at most polynomially with rate $p = 1$.

PROOF. On the event $\{S_n \geq na\}$,

$$L_g(S_n) \leq \left(\int_{a_1}^{a_2} \exp[\theta(t) n(a - \Lambda(\theta(t)))] g(t)\, dt \right)^{-1}.$$

Because $\Lambda^*(t) = \theta(t)t - \Lambda(\theta(t))$, and $\theta(t) = \Lambda^{*\prime}(t)$, this upper bound equals

$$\exp(-n\Lambda^*(a)) \left(\int_{a_1}^{a_2} \exp(-n(\Lambda^*(a) - \Lambda^{*\prime}(t)(a-t) - \Lambda^*(t)) g(t)\, dt \right)^{-1}.$$

Note that $\Lambda^*(a) - \Lambda^{*\prime}(t)(a-t) - \Lambda^*(t)$ is minimized at $t = a$. Using Laplace's approximation (see, e.g., Olver [12, pp. 80–81]),

$$\int_{a_1}^{a_2} \exp(-n(\Lambda^*(a) - \Lambda^{*\prime}(t)(a-t) - \Lambda^*(t)) g(t)\, dt \sim \sqrt{\frac{2\pi}{n\Lambda''(\theta_a)}} g(a) \tag{34}$$

if $a \in (a_1, a_2)$. Thus, for $\epsilon > 0$ and n large enough, on $\{S_n \geq na\}$,

$$L_g(S_n) \leq \exp(-n\Lambda^*(a)) \sqrt{\frac{n\Lambda''(\theta(a))}{2\pi}} \frac{1}{g(a)} (1 + \epsilon). \tag{35}$$

Now $E_{P_g}[L_g(S_n)^2 I(S_n/n \geq a)] = E_P[L_g(S_n) I(S_n/n \geq a)]$. Hence, for n large enough,

$$E_{P_g}[L_g(S_n)^2 I(S_n/n \geq a)] \leq \exp(-n\Lambda^*(a)) \sqrt{\frac{n\Lambda''(\theta_a)}{2\pi}} \frac{1}{g(a)} (1 + \epsilon) P(S_n/n \geq a).$$

Using the sharp asymptotic for $P(S_n/n \geq a)$ given in (8), and noting that ϵ is arbitrary, it follows that

$$\limsup_{n\to\infty} \frac{1}{n} \frac{E_{P_g}[L_g(S_n)^2 I(S_n/n \geq a)]}{P(S_n/n \geq a)^2} \leq \frac{\theta_a \Lambda''(\theta_a)}{g(a)}.$$

Because $\theta_a = \Lambda^{*\prime}(a)$ and $\Lambda''(\theta_a) = 1/\Lambda^{*\prime\prime}(a)$, the result follows. □

REMARK 4.1. In Theorem 4.1, if a is on the boundary of $[a_1, a_2]$, the analysis remains unchanged except that in (34) the asymptotic has to be divided by two to get the correct value (see, e.g., Olver [12, p. 81]). Hence, in this setting

$$\limsup_{n\to\infty} \frac{1}{n} \frac{E_{P_g}[L_g(S_n)^2 I(S_n/n \geq a)]}{P(S_n/n \geq a)^2} \leq 2 \frac{\Lambda^{*\prime}(a)}{\Lambda^{*\prime\prime}(a) g(a)}.$$

In particular, it follows that if $\mathcal{A} = [a_1, a_2] \subseteq \mathcal{H}^o$, $a_1 > \mu$,

$$\sup_{a\in\mathcal{A}} \limsup_{n\to\infty} \frac{1}{n} \frac{E_{P_g}[L_g(S_n)^2 I(S_n/n \geq a)]}{P(S_n/n \geq a)^2} < \infty.$$

REMARK 4.2. A difficulty with implementing a continuous mixture importance sampling distribution P_g may be that the likelihood $L_g(S_n)$ may not be computable in closed form and the numerical methods to compute $L_g(S_n)$ may be computationally expensive. Hence, in practice a discrete mixture may often be easier to implement compared to a continuous mixture.

4.2. Choice of density function. Theorem 4.1 and Remark 4.1 suggest that the relative variance of the estimator of $P(S_n/n \geq a)$ for each $a \in \mathcal{A}$ is closely related to the quantity $\Lambda^{*\prime}(a)/(\Lambda^{*\prime\prime}(a)g(a))$. In particular, all else being equal, if $g(a)$ is reduced by a factor α, then the associated variance roughly increases by a factor α. This is intuitive because $g(a)$ is a rough measure of the proportion of samples generated in the neighborhood of a under P_g (recall that under the probability measure $P_{\theta(a)}$, the mean of S_n/n equals a), and (35) suggests that the simulation output on the set $\{S_n/n \geq a\}$ depends on $g(\cdot)$ primarily through $g(a)$. For instance, consider the case where $a \in \mathcal{A} = [a_1, a_1 + \delta]$ and $g(\cdot)$ is the density of the uniform distribution on this interval. Now if the interval width is doubled so that $\mathcal{A} = [a_1, a_1 + 2\delta]$, and if $g(\cdot)$ is again chosen to be a density of the uniform distribution along the new interval, then it is reasonable to expect that the estimator of $P(S_n/n \geq a)$ will have about double the variance in the new settings.

From an implementation viewpoint, it may be desirable to select g to estimate all $P(S_n/n \geq a)$, $a \in \mathcal{A}$, with the same asymptotic relative variance. If this holds, we may continue to generate samples until the sample relative variance at one point becomes small, and this ensures that the relative variance at other points is not much different. Theorem 4.1 suggests that this may be achieved by selecting $g(a)$ proportional to $\Lambda^{*\prime}(a)/\Lambda^{*\prime\prime}(a)$ or, equivalently, $\theta(a)/\theta'(a)$. Suppose that $\mathcal{A} = [a_1, a_2] \subseteq \mathcal{H}^o$, $a_1 > \mu$. In this case

$$\int_{a_1}^{a_2} \frac{\theta(a)}{\theta'(a)}\, da < \infty, \tag{36}$$

so such a g is feasible.

We now evaluate such a g for some simple cases:

EXAMPLE 4.1. If X_i has a Bernoulli distribution with mean p, then its log-moment generating function is

$$\Lambda(\theta) = \log(\exp(\theta)p + (1-p))$$

and

$$\Lambda'(\theta) = \frac{\exp(\theta)p}{\exp(\theta)p + (1-p)}.$$

It is easily seen that

$$\theta(t) = \log\left[\frac{t}{1-t}\frac{1-p}{p}\right]$$

and $\theta'(t) = 1/t + 1/(1-t) = 1/[t(1-t)]$. Thus, $g(t)$ equals

$$ct(1-t)\log\left[\frac{t}{1-t}\frac{1-p}{p}\right],$$

where c is a normalization constant so that $\int_{a_1}^{a_2} g(t)\, dt = 1$. It can be seen that this function is maximized at the solution of

$$\frac{t}{1-t}\exp[-1/(2t-1)] = \frac{p}{1-p}$$

that exceeds p. The likelihood ratio in (33) specialized to this setting equals

$$L_g(S_n) = \left(c\int_{a_1}^{a_2} \left(\frac{t}{p}\right)^{S_n} \left(\frac{1-t}{1-p}\right)^{n-S_n} t(1-t)\log\left[\frac{t}{1-t}\frac{1-p}{p}\right] dt\right)^{-1}.$$

EXAMPLE 4.2. If X_i has a Gaussian distribution with mean zero and variance one, then its log-moment generating function is $\Lambda(\theta) = \theta^2/2$ and $\Lambda'(\theta) = \theta$. Also, $\theta(t) = t$ and $\theta'(t) = 1$ and $g(t)$ equals $2t/(a_2^2 - a_1^2)$. This suggests that more mass should be given to larger values of t in the interval $[a_1, a_2]$. The likelihood ratio in this setting equals

$$L_g(S_n) = \left(\frac{2}{a_2^2 - a_1^2} \int_{a_1}^{a_2} \exp[tS_n - nt^2/2]t \, dt \right)^{-1}.$$

EXAMPLE 4.3. If X_i has a gamma distribution with log-moment generating function

$$\Lambda(\theta) = -\alpha \log(1 - \theta\beta),$$

then its mean equals $\alpha\beta$. Now

$$\Lambda'(\theta) = \frac{\alpha\beta}{1 - \theta\beta}$$

so that $\theta(t) = (1/\beta)(1 - \alpha\beta/t)$ and $\theta'(t) = \alpha/t^2$. Then, $g(t)$ equals

$$ct\left(\frac{t}{\alpha\beta} - 1\right),$$

where

$$c = \left(\frac{a_2^3 - a_1^3}{3\alpha\beta} - \frac{a_2^2 - a_1^2}{2} \right)^{-1}.$$

Recall that $t > \alpha\beta$. The likelihood ratio in this setting equals

$$L_g(S_n) = \left(c \int_{a_1}^{a_2} \exp\left[\frac{1}{\beta}\left(1 - \frac{\alpha\beta}{t}\right) S_n \right] \left(\frac{\alpha\beta}{t}\right)^{\alpha n} t\left(\frac{t}{\alpha\beta} - 1\right) dt \right)^{-1}.$$

5. Concluding remarks. In this paper, we have considered the problem of simultaneous estimation of the probabilities of multiple rare events. In the setting of tail probabilities associated with a random walk, we have shown that the standard importance sampling estimator that yields asymptotically efficient estimates for one point on the distribution fails to do so for all other points. To address this problem, we have examined mixtures of exponentially twisted distributions. We have identified the optimal finite mixture and shown that asymptotic efficiency is achieved uniformly with either a continuous mixture or a finite mixture with an increasing number of components. Although our analysis is restricted to the random walk setting, we expect that similar techniques could prove useful in other rare-event simulation problems. Similar ideas should also prove useful in going beyond the estimation of multiple probabilities to the estimation of functions of tail distributions, such as quantiles or tail conditional expectations.

Acknowledgment. This work is supported, in part, by NSF Grants DMI0300044 and DMS0410234.

References

[1] Arsham, H., A. Fuerverger, D. L. McLeish, J. Kreimer, R. Y. Rubinstein. 1989. Sensitivity analysis and the "what if" problem in simulation analysis. *Math. Comput. Model.* **12** 193–219.
[2] Asmussen, S. 2000. *Ruin Probabilities*. World Scientific, London.
[3] Bahadur, R. R., R. R. Rao. 1960. On deviations of the sample mean. *Ann. Math. Statist.* **31** 1015–1027.
[4] Bucklew, J. A. 2004. *Introduction to Rare Event Simulation*. Springer-Verlag, New York.
[5] Bucklew, J. A., P. Ney, J. S. Sadowsky. 1990. Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains. *J. Appl. Probab.* **27** 44–59.
[6] Chang, C. S., P. Heidelberger, S. Juneja, P. Shahabuddin. 1994. Effective bandwidth and fast simulation of ATM intree networks. *Performance Eval.* **20** 45–65.
[7] Dembo, A., O. Zeitouni. 1998. *Large Deviations Techniques and Applications*, 2nd ed. Springer, New York.
[8] Glasserman, P., J. Li. 2005. Importance sampling for portfolio credit risk. *Management Sci.* **51** 1643–1656.
[9] Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Trans. Model. Comput. Simulation* **5**(1) 43–85.
[10] Heidelberger, P., V. F. Nicola, P. Shahabuddin. 1992. Simultaneous and efficient simulation of highly dependable systems with different underlying distributions. J. J. Swaim, D. Goldsman, R. C. Crain, J. R. Wilson, eds. *Proc. 1992 Winter Simulation Conf.*, IEEE Press, Piscataway, NJ.
[11] Juneja, S., P. Shahabuddin. 2006. Rare event simulation techniques. S. Henderson, B. Nelson, eds. *Handbooks in Operations Research and Management Science*, Vol. 13, *Simulation*. Elsevier North-Holland, Amsterdam, 291–350.
[12] Olver, F. W. J. 1974. *Introduction to Asymptotics and Special Functions*. Academic Press, New York.

[13] Sadowsky, J. S., J. A. Bucklew. 1990. On large deviation theory and asymptotically efficient Monte Carlo estimation. *IEEE Trans. Inform. Theory* **36**(3) 579–588.
[14] Sadowsky, J. S. 1991. Large deviations theory and efficient simulation of excessive backlogs in $GI/GI/m$ queue. *IEEE Trans. Automat. Control* **36** 1383–1394.
[15] Shahabuddin, P. 1994. Importance sampling for the simulation of highly reliable Markovian systems. *Management Sci.* **40** 333–352.
[16] Siegmund, D. 1976. Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.* **4** 673–684.