*Proceedings of the 2005 Winter Simulation Conference*
*M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.*

# IMPORTANCE SAMPLING SIMULATION IN THE PRESENCE OF HEAVY TAILS

Achal Bassamboo

Kellogg School of Management
Northwestern University
Evanston, IL U.S.A.

Sandeep Juneja

School of Technology and Computer Science
Tata Institute of Fundamental Research,
Mumbai, India

Assaf Zeevi

Graduate School of Business,
Columbia University,
New York, NY U.S.A.

## ABSTRACT

We consider importance sampling simulation for estimating rare event probabilities in the presence of heavy-tailed distributions that have polynomial-like tails. In particular, we prove the following negative result: there does not exist an asymptotically optimal *state-independent* change-of-measure for estimating the probability that a random walk (respectively, queue length for a single server queue) exceeds a "high" threshold before going below zero (respectively, becoming empty). Furthermore, we derive explicit bounds on the best asymptotic variance reduction achieved by importance sampling relative to naïve simulation. We illustrate through a simple numerical example that a "good" *state-dependent* change-of-measure may be developed based on an approximation of the zero-variance measure.

## 1 Introduction

Importance sampling (IS) simulation has proven to be an extremely successful method in efficiently estimating certain rare events associated with *light-tailed* random variables; see, e.g., Sadowsky (1991) and Heidelberger (1995) for queueing and reliability applications, and Glasserman (2003) for applications in financial engineering. (Roughly speaking, a random variable is said to be light-tailed if the tail of the distribution decays at least exponentially fast.) The main idea of IS algorithms is to perform a *change-of-measure*, then estimate the rare event in question by generating iid copies of the underlying random variables according to this new distribution. A good IS distribution not only assigns high probability to the most likely paths to the rare events but *equally importantly* it does not significantly reduce the probability of other less likely paths.

Recently, *heavy-tailed* distributions have become increasingly important in explaining rare event related phenomena in many fields including data networks and teletraffic models (see, e.g., Resnick (1997)), and insurance and risk management (cf. Embrechts, Klppelberg & Mikosch (1997)). Unlike the light-tailed case, designing efficient IS simulation techniques in the presence of heavy-tailed random variables has proven to be quite challenging. This is mainly due to the fact that the manner in which rare events occur is quite different than that encountered in the light-tailed context (see, Asmussen (1998) for further discussion).

In this paper we highlight a fundamental difficulty in applying IS techniques in the presence of heavy-tailed random variables. For a broad class of such distributions having polynomial-like tails, we prove that if the constituent random variables are independent under an IS change-of-measure then it cannot achieve *asymptotic optimality*. (Roughly speaking, a change-of-measure is said to be asymptotically optimal if it asymptotically achieves zero variance on a logarithmic scale; a precise definition is given in Section 2.) In particular, we give explicit asymptotic bounds on the level of improvement that state-independent IS can achieve vis-a-vis naïve simulation. These results are derived for the following two rare events.

i) A negative drift random walk (RW) $S_n = \sum_{i=1}^{n} X_i$ exceeding a large threshold before taking on a negative value (see Theorem 1), as well as $\max\{S_n : n = 1, 2, \ldots\}$ exceeding a large threshold.

ii) A stable GI/GI/1 queue exceeding a large threshold within a busy cycle (see Theorem 2). This analysis relies on asymptotes for the maximum of the queue length process (see Proposition 1).

The above probabilities are particularly important in estimating steady-state performance measures related to waiting times and queue lengths in single-server queues, when the regenerative ratio representation is exploited for estimation (see, e.g., Heidelberger (1995)). Our negative results motivate the development of state-dependent IS techniques (see, e.g., Kollman, Baggerly, Cox & Picard (1999), and Blanchet & Glynn (2005)). In particular, for the probabilities that we consider

the zero variance measure has a straightforward "state-dependent" representation. In the random walk setting this involves generating each increment $X_i$ using a distribution that depends on the position of the RW prior to that, *i.e.*, the distribution of $X_i$ depends on $S_{i-1} = \sum_{j=1}^{i-1} X_j$. For a simple example involving a slotted time queue, we illustrate numerically how one can exploit approximations to the zero-variance measure (see Proposition 2) to develop state-dependent IS schemes that perform reasonably well.

**Related literature.** The first algorithm for efficient simulation in the heavy-tailed context was given in Asmussen & Binswanger (1997) using conditional Monte Carlo. Both Asmussen, Binswanger & Hojgaard (2000) and Juneja & Shahabuddin (2002) develop successful IS techniques to estimate level crossing probabilities of the form $P(\max_n S_n > u)$, for random walks with heavy tails, by relying on the ladder height representation of this probability. However, the ladder height representation is useful for a restricted class of random walks (where each $X_i$ is a difference of a heavy tailed random variable and an exponentially distributed random variable). The work in Boots & Shahabuddin (2001) also considers the level crossing problem and obtains positive results for IS simulation in the presence of Weibull-tails. They avoid the inevitable variance build-up by truncating the generated paths. However, even with truncation they observe poor results when the associated random variables have polynomial tails. Recently, Blanchet & Glynn (2005) described an asymptotically optimal state-dependent change-of-measure for the probability that the maximum of a negative drift random walk exceeding a large threshold.

In terms of negative results, Asmussen, Kroese & Rubinstein (2004) show that performing a change in parameters within the family of Weibull or Pareto distributions does not result in an asymptotically optimal IS scheme in the random-walk or in the single server queue example. Our paper provides further evidence that any state-independent change-of-measure (not restricted to just parameter changes in the original distribution) will not lead to efficient IS simulation algorithms. We also explicitly bound the loss of efficiency that results from restricting use to iid IS distributions.

**The remainder of this paper.** In Section 2, we briefly describe IS and the notion of asymptotic optimality. Section 3 describes the main results of the paper. In Section 4 we illustrate empirically the performance of a state-dependent change-of-measure for a simple discrete time queue. We conclude in Section 5 with some general observations related to this paper. All proofs are collected in Appendix A.

## 2 Importance Sampling and Asymptotic Optimality

### 2.1 Two rare events

**Random walk.** Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random walk $S_n = \sum_{m=1}^{n} X_m$, $S_0 = 0$ where $X_1, X_2, ...$ are iid copies of $X$. We assume that $\mathbb{E}X < 0$, and we denote the cumulative distribution function of $X$ by $F$. Define $\tau$ to be the time at which the random walk first goes below zero, i.e.,

$$\tau = \inf\{n \geq 1 : S_n < 0\}.$$

Let $\zeta = \mathbb{E}\tau$, and $M_n = \max_{0 \leq m \leq n} S_m$. The probability of interest is either $\gamma_u = \mathbb{P}(M_\tau > u)$ or the probability that the "all-time-max" of the random walk exceeds level $u$, viz., $\mathbb{P}(M_\infty > u)$. To fix ideas, let us focus on the former probability. To estimate this probability by naïve simulation, we generate $m$ iid samples of the function $\mathbb{I}_{\{M_\tau > u\}}$ and average over them to get an unbiased estimate $\hat{\gamma}_u^m$. The relative error of this estimator (defined as the ratio of standard deviation and mean) is given by $\sqrt{\frac{(1-\gamma_u)}{m\gamma_u}}$. Since $\gamma_u \to 0$ as $u \to \infty$, the number of simulation runs must increase without bound in order to have fixed small relative error as $u$ becomes large.

Consider another probability distribution $\widetilde{\mathbb{P}}$ on the same sample space such that the sequence $\{X_1, X_2, ...\}$ is iid under $\widetilde{\mathbb{P}}$ with marginal distribution $\widetilde{F}$, and $F$ is absolutely continuous w.r.t. $\widetilde{F}$. Let $T_u = \inf\{n : S_n \geq u\}$. Define

$$Z_u = L_u \mathbb{I}_{\{M_\tau > u\}}, \tag{1}$$

$$\text{where} \qquad L_u = \prod_{i=1}^{\min\{\tau, T_u\}} \frac{dF(X_i)}{d\widetilde{F}(X_i)},$$

and let $\widetilde{\mathbb{E}}[\cdot]$ be the expectation operator under $\widetilde{\mathbb{P}}$. Then, using Wald's likelihood ratio identity (see Siegmund (1985, Proposition 2.24)), we have that $Z_u$ under measure $\widetilde{\mathbb{P}}$ is an unbiased estimator of the probability $\mathbb{P}(M_\tau > u)$. Thus, we can generate iid samples of $Z_u$ under the measure $\widetilde{\mathbb{P}}$, the average of these would be an unbiased estimate of $\gamma_u$. We refer to $\widetilde{\mathbb{P}}$ as the IS change-of-measure and $L_u$ as the likelihood ratio. By choosing the IS change-of-measure appropriately, we can substantially reduce the variance of this estimator.

Note that a similar analysis can be carried out to get an estimator when the sequence $\{X_1, X_2, ...\}$ is not iid under $\widetilde{\mathbb{P}}$. The likelihood ratio $L_u$ in that case can be expressed as the Radon-Nikodyn derivative of the original measure w.r.t. the IS measure restricted to the appropriate stopping time. (A similar construction

with slight modification applies in the case of the all-time-max problem; we omit details.)

**Queue length process.** The second rare event studied in this paper is the buffer overflow during a busy cycle. Consider a GI/GI/1 queue, and let the inter-arrival and service times have finite means $\lambda^{-1}$ and $\mu^{-1}$, respectively. Let $Q(t)$ represent the queue length at time $t$ under FCFS (first come first serve) service discipline. Assume that the busy cycle starts at time $t = 0$, *i.e*, $Q(0) = 1$, and let $\tau$ denote the end of the busy cycle, namely

$$\tau = \inf\{t \geq 0 : \ Q(t^-) > 0, Q(t) = 0\}.$$

Let the cumulative distribution of inter-arrival times and service times be $F$ and $G$, respectively. Let $S_i$ be the service time of the $i^{th}$ customer and $A_i$ be the inter-arrival time for the $(i+1)^{th}$ customer. The probability of interest is $\gamma_u = \mathbb{P}(\max_{0\leq t\leq \tau} Q(t) \geq u)$. Again we note that $\gamma_u \to 0$ as $u \to \infty$; to estimate this probability efficiently we can use IS.

Let the number of arrivals until the queue length exceeds level $u$ be

$$M = \inf\left\{n \geq 1 : \sum_{i=1}^{n} A_i < \sum_{i=1}^{n-u+2} S_i\right\}.$$

Let $N(t)$ represent the number of arrivals up until time $t$. Then $N(\tau)$ is the number of customers arriving during a busy period. Let $\widetilde{F}$ and $\widetilde{G}$ be the cumulative IS distributions of inter-arrival and service times, respectively. Then, again using Wald's likelihood ratio identity, $Z_u$ under the measure $\widetilde{\mathbb{P}}$ is an unbiased estimator for the probability $\mathbb{P}(\max_{0\leq t\leq \tau} Q(t) > u)$, where

$$Z_u = L_u \mathbb{I}_{\{M\leq N(\tau)\}}, \tag{2}$$

$$\text{and} \qquad L_u = \prod_{i=1}^{M} \frac{dF(A_i)}{d\widetilde{F}(A_i)} \prod_{j=1}^{M-u+2} \frac{dG(S_j)}{d\widetilde{G}(S_j)}.$$

### 2.2 Asymptotic Optimality

Consider a sequence of rare-events indexed by a parameter $u$. Let $\mathbb{I}_u$ be the indicator of this rare event, and suppose $\mathbb{E}[\mathbb{I}_u] \to 0$ as $u \to \infty$ (e.g., for the first rare event defined above, $\mathbb{I}_u = \mathbb{I}_{\{M_\tau > u\}}$). Let $\widetilde{\mathbb{P}}$ be an IS distribution and $L$ be the corresponding likelihood ratio. Put $Z_u = L\mathbb{I}_u$.

**Definition 1 (asymptotic optimality)** *A sequence of IS estimators is said to asymptotically optimal if*

$$\frac{\log \widetilde{\mathbb{E}}[Z_u^2]}{\log \widetilde{\mathbb{E}}[Z_u]} \to 2 \quad as \quad u \to \infty. \tag{3}$$

Note that $\widetilde{\mathbb{E}}[Z_u^2] \geq (\widetilde{\mathbb{E}}[Z_u])^2$, therefore for any sequence of IS estimators we have

$$\limsup_{u\to\infty} \frac{\log \widetilde{\mathbb{E}}[Z_u^2]}{\log \widetilde{\mathbb{E}}[Z_u]} \leq 2.$$

(Note that $\log \widetilde{\mathbb{E}}[Z_u] < 0$.) Thus, loosely speaking, asymptotic optimality implies minimal variance on logarithmic scale.

## 3 Main Results

### 3.1 Random walk

Consider the random walk defined in Section 2.1. We assume that the distribution of $X$ satisfies

$$\begin{aligned} &\frac{\log \mathbb{P}(X > x)}{\log x} \to -\alpha, \\ \text{and} \quad &\frac{\log \mathbb{P}(X < -x)}{\log x} \to -\beta, \end{aligned} \tag{4}$$

as $x \to \infty$, where $\alpha \in (1, \infty)$ and $\beta \in (1, \infty]$. Further, we assume that $\mathbb{P}(X > x) \sim 1 - B(x)$ as $x \to \infty$, for some distribution $B$ on $(0, \infty)$ which is *subexponential*, that is, it satisfies

$$\limsup_{x\to\infty} \frac{1 - (B * B)(x)}{1 - B(x)} \leq 2,$$

where '$*$' denotes the convolution operator (cf. Embrechts et al. (1997)). We write $f(u) \sim g(u)$ as $u \to \infty$ if $\frac{f(u)}{g(u)} \to 1$ as $u \to \infty$. Thus, distributions with regularly varying tails are a subset of the class of distributions satisfying our assumptions. (Regularly varying distributions have $1 - F(x) = \mathcal{L}(x)/x^\alpha$, where $\alpha > 1$ and $\mathcal{L}(x)$ is slowly varying; for further discussion see Embrechts et al. (1997, Appendix A.3).) Note that (4) allows the tail behavior on the negative side to be lighter than polynomial as $\beta = \infty$ is permitted. We denote the cumulative distribution function of $X$ by $F$. From Asmussen (1998) it follows that

$$\mathbb{P}(M_\tau > u) \sim \zeta \mathbb{P}(X > u) \text{ as } u \to \infty, \tag{5}$$

where $\zeta$ is the expected time at which the random walk goes below zero. In the case of the all-time-max the counterpart of (5) is given in Theorem 3 in the appendix.

Consider the IS probability distribution $\widetilde{\mathbb{P}}$ such that the sequence $\{X_1, X_2, ...\}$ is iid under $\widetilde{\mathbb{P}}$ with marginal distribution $\widetilde{F}$, and $F$ is absolutely continuous w.r.t. $\widetilde{F}$. Let $\mathcal{P}$ be the collection of all such probability distributions on the sample space $(\Omega, \mathcal{F})$. Let $Z_u$ be the estimator defined in (1). Thus, $\widetilde{\mathbb{E}}[Z_u]$ is an unbiased estimator of $\mathbb{P}(M_\tau > u)$. We then have the following result.

**Theorem 1** *For any* $\widetilde{\mathbb{P}} \in \mathcal{P}$

$$\limsup_{u\to\infty} \frac{\log \widetilde{\mathbb{E}}[Z_u^2]}{-\alpha \log u} \leq 2 - \frac{\min(\alpha,\beta)}{\alpha(1+\min(\alpha,\beta))},$$

*where* $\alpha$ *and* $\beta$ *are defined in (4).*

**Intuition and proof sketch.** The proof follows by contradiction. We consider two disjoint subsets $\mathcal{B}$ and $\mathcal{C}$ of the "rare set" $\mathcal{A} = \{\omega : M_\tau > u\}$ and use the fact that $\widetilde{\mathbb{E}}[L_u^2 \mathbb{I}_{\{\mathcal{A}\}}] \geq \widetilde{\mathbb{E}}[L_u^2 \mathbb{I}_{\{\mathcal{B}\}}] + \widetilde{\mathbb{E}}[L_u^2 \mathbb{I}_{\{\mathcal{C}\}}]$. The two sets are as follows.

1) The subset $\mathcal{B}$ consists of sample paths where the first random variable is "large" and causes the random walk to immediately exceed level $u$.

2) The subset $\mathcal{C}$ which consists of sample paths where the $X_i$'s are of order $u^\gamma$ for $i = 2, \ldots, \lfloor u^{1-\gamma} \rfloor$ followed by one "big" jump.

Assuming that the limit in the above theorem is violated, we consider the sample paths in set $\mathcal{B}$ to obtain a lower bound on the probability that $X$ exceeds $u$ under the IS distribution $\widetilde{F}$. The above, in turn, restricts the mass that can be allocated below level $u$. We then consider the subset $\mathcal{C}$, and by selecting the parameter $\gamma$ and the value of $X_1$ judiciously, we show that the second moment on the set $\mathcal{C}$ is infinite. (See Bassamboo, Juneja & Zeevi (2005) for details of the rigorous proof.)

**Extension to all-time-max problem.** The non-asymptotic optimality of the state independent change-of-measure can be analogously seen for the all-time-max problem. Again, we note that if the performance of the proposed importance sampling algorithm is close to asymptotically optimal, it must assign significant probability to the set $\{X > u\}$ (this can be seen by considering the contribution of $\{X_1 > u\}$ to the second moment). This provides an upper bound on the probability mass assigned to the set $X \in (-\log u, u^\gamma)$, for any $\gamma < 1$. Now consider a set of paths where the first jump is negative, taking values of order $-u^\beta$ for $\beta > 1$, the remaining $u^{\beta-\gamma}$ increments take values between $(-\log u, u^\gamma)$, and the last increment ensures that the threshold $u$ is crossed. Along these paths an upper bound on efficiency improvement may be constructed by appropriately selecting $\beta$ and $\gamma$.

### 3.2 Queue length process

Consider a GI/GI/1 queue described in Section 2.1 with service times being iid copies of $S$ and inter-arrival times being iid copies of $A$. Put $\Lambda(x) := -\log \mathbb{P}(S > x) = -\log(1 - G(x))$. Assume that

$$\frac{\Lambda(x)}{\log x} \to \alpha \text{ as } x \to \infty, \tag{6}$$

where $\alpha \in (1, \infty)$, and $(S - A)$ has a subexponential distribution. We then have the following logarithmic asymptotics for the buffer overflow probability in a busy cycle.

**Proposition 1** *Let assumption (6) hold. Then,*

$$\lim_{u\to\infty} \frac{\log \mathbb{P}(\max_{0\leq t\leq \tau} Q(t) > u)}{\log u} = -\alpha.$$

Recall that $\widetilde{F}$ and $\widetilde{G}$ are the cumulative IS distribution of inter-arrival and service times, respectively, and an unbiased estimator for the probability $\mathbb{P}(\max_{0\leq t\leq \tau} Q(t) > u)$ is $\widetilde{\mathbb{E}}[Z_u]$ where $Z_u$ is as defined in (2). Let $\widetilde{\mathbb{P}}$ be the product measure generated by $(\widetilde{F}, \widetilde{G})$, and let $\mathcal{D}$ be the collection of all such measures.

**Theorem 2** *For any* $\widetilde{\mathbb{P}} \in \mathcal{D}$

$$\limsup_{u\to\infty} \frac{\log \widetilde{\mathbb{E}}[Z_u^2]}{-\alpha \log u} \leq 2 - \frac{1}{1+\alpha}.$$

**Intuition and proof sketch.** The proof of the above theorem is similar to proof of Theorem 1. We again consider two sets and arrive at a contradiction. The sets in this case are given as follows.

1) The first set of sample paths are those for which $\{S_1(\omega) > 2u\lambda^{-1}\}$ and $\{\sum_{i=1}^{u} A_i < 2u\lambda^{-1}\}$.

2) The second set of sample paths are defined as follows.

   (a) The first service time $S_1 \in [2u^{1-\gamma}\lambda^{-1}, 3u^{1-\gamma}\lambda^{-1}]$.

   (b) The sum of the first $\lfloor u^{1-\gamma} \rfloor$ inter-arrival times is less than $2u^{1-\gamma}\lambda^{-1}$, i.e., $\sum_{i=1}^{\lfloor u^{1-\gamma} \rfloor} A_i \leq 2u^{1-\gamma}\lambda^{-1}$. This ensures that by the end of service of the first customer at least $\lfloor u^{1-\gamma} \rfloor$ customers are in the queue.

   (c) The next $\lfloor u^{1-\gamma} \rfloor - 1$ services lie in the interval $[0, 0.5u^\gamma\lambda^{-1}]$. This ensures that at most $0.5u\lambda^{-1}$ time has elapsed before the beginning of service of customer $\lfloor u^{1-\gamma} \rfloor$.

   (d) The service time for customer $\lfloor u^{1-\gamma} \rfloor$ exceeds $2u\lambda^{-1}$.

   (e) The next $\lfloor 0.6u \rfloor$ arrivals are such that $0.5u\lambda^{-1} \leq \sum_{i=\lfloor u^{1-\gamma} \rfloor+1}^{\lfloor u^{1-\gamma} \rfloor+\lfloor 0.6u \rfloor} A_i \leq 0.75u\lambda^{-1}$. This ensures that the buffer does not overflow before the beginning of service of customer $\lfloor u^{1-\gamma} \rfloor$.

   (f) The next $\lfloor 0.4u \rfloor$ arrivals are such that $0.3u\lambda^{-1} \leq \sum_{i=\lfloor u^{1-\gamma} \rfloor+\lfloor 0.6u \rfloor}^{\lfloor u^{1-\gamma} \rfloor+\lfloor u \rfloor} A_i \leq 0.75u\lambda^{-1}$. This ensures that the buffer overflows during the service of customer $\lfloor u^{1-\gamma} \rfloor$.

(Here $\gamma$ is a constant chosen appropriately.) First set of sample paths are used to lower bound the probability allocated to tails of service time distribution. This results in the fact that the services in condition $2(c)$ are assigned sufficiently low probability under the new measure and thus the second moment of the estimator builds up along such realizations. The remaining conditions for the set defined in 2) ensure that the buffer overflows for the paths in this set. (See Bassamboo et al. (2005) for details of the rigorous proof.)

## 4 State Dependent Change-of-Measure

In this section we outline a general principle that can guide the construction of "good" state-dependent changes of measure, and illustrate it via a simple example. The main idea is to use a suitable approximation for the zero variance measure. In particular, for probabilities involving random walks hitting a rare set, as is the case for the probabilities studied in this paper, the zero variance measure has a simple Markovian representation.

**Preliminaries.** Consider a discrete time queuing system where the probability of interest is buffer overflow during a busy cycle. Specifically, the time axis is divided into fixed-length time intervals called *slots* and each service requires one slot. During time slot $n$, $X_n$ customers arrive where $\{X_n : n \geq 1\}$ is a sequence of iid random variables with $\mathbb{E}[X_1] < 1$. Let $Q_n$ represent the number of customers in the system at the beginning of time slot $n$. We then have the following recursion

$$Q_n = \max\{Q_{n-1} - 1, 0\} + X_n.$$

Let $Q_0 = 1$, $\tau_0 = \inf\{m > 0 : Q_m = 0\}$ and $\tau_u = \inf\{m > 0 : Q_m \geq u\}$. The rare event of interest is $\{\tau_u < \tau_0\}$ when $u$ is large. We assume that $X_1$ is a regularly varying distribution with parameter $\alpha$, i.e.,

$$1 - F(x) = \mathbb{P}(X_1 > x) = \mathcal{L}(x)/x^{\alpha},$$

for all $x$, where $\alpha \in (1, \infty)$, $\mathcal{L}(x)$ is slowly varying, and $F(\cdot)$ is the cumulative density function of $X$. Using results from Section 3, we know that any state-independent change to the distribution of the $X_i$'s cannot yield an asymptotically optimal IS estimator. We also know that there always exists a change-of-measure (which may be state-dependent) that has zero variance, (cf. Ahamed, Borkar & Juneja (2005)) and in this setting this change-of-measure has a Markovian structure. In particular, let $J_y(u) = \mathbb{P}(\tau_u < \tau_0 | Q(0) = y)$ for all $y = \{0, 1, \ldots\}$. Then, under the zero variance measure $X_n$ has distribution

$$\widetilde{\mathbb{P}}(X_n = x | Q_{n-1} = y) = \frac{\mathbb{P}(X_n = x) J_{x+y-1}(u)}{J_y(u)} \qquad (7)$$

for all $x \in \{0, 1, \ldots\}$ and $n = 1, 2, \ldots$. We shall now develop asymptotics for $J_y(u)$ and use them to construct a "good" state-dependent IS change-of-measure.

**Proposition 2** *For all $\beta \in (0, 1)$*

$$J_{\lfloor \beta u \rfloor}(u) \sim \mathbb{E}[N] \left[ \int_{x=(1-\beta)u}^{u} (1 - F(x)) dx \right] \text{ as } u \to \infty, \quad (8)$$

*where $N$ is the number of arrivals during a busy period.*

As is evident from the proof given in the Appendix, the above proposition may be extended to continuous state space under mild regularity conditions.

**Description of the numerical example.** For the purpose of our numerical study, we consider an M/GI/1 queue whose arrival stream is Poisson with rate $\lambda$ and service times are iid copies of $S$ having Pareto distribution with parameter $\alpha \in (1, \infty)$, *i.e.*,

$$\mathbb{P}(S \geq x) = \begin{cases} x^{-\alpha} & \text{if } x \geq 1 \\ 1 & \text{otherwise.} \end{cases}$$

The embedded Markov chain in this system evolves as the discrete-time queue described above where $X_1 \stackrel{\mathrm{d}}{=} Poisson(\lambda S)$. The (state-dependent) IS distribution we propose is

$$\widetilde{\mathbb{P}}(X_n = x | Q_{n-1} = y) = \frac{g(x, y)}{\sum_{x=0}^{\infty} g(x, y)},$$

$$\text{where} \qquad g(x, y) = \mathbb{P}(X_1 = x) \left[ \sum_{x'=u-x-y+2}^{u+1} \mathbb{P}(X_1 \geq x') \right].$$

We obtain the above change-of-measure by substituting the asymptotes from Proposition 2 in the zero variance measure given in (7). Note that it is easy to compute $g(x, y)$ in this simple setting, and it can be expressed as a product of a function of $x$ and a function of $x + y$. We simulate the results for the following cases: buffer levels $u = 100$ and 1000; tail parameter values $\alpha = 2, 9$ and 19; and traffic intensities $\rho = 0.3, 0.5$ and 0.8. (The traffic intensity $\rho$ equals $\lambda\alpha/(\alpha - 1)$.) The number of simulation runs in all cases is taken to be $500,000$. To test the accuracy of the simulation results, we also calculate the buffer overflow probabilities using first step analysis.

**Simulation results.** Results in Table 1 illustrate the following points. First, the accuracy of the proposed IS method decreases as the traffic intensity increases, and/or the tail becomes "lighter." Second, accuracy for the problem involving buffer level 1000 is better than the case of buffer level 100, in accordance with the fact that we are using a "large buffer" asymptotic approximation to the zero variance measure. Finally, the relative error on logarithmic scale is quite close to the

best possible value of 2, hence we anticipate that our proposed IS scheme might be asymptotically optimal. The rigorous derivations of such results and their generalizations to continuous state space is left for future work.

**Discussion.** In a recent work, Ahamed et al. (2005) propose stochastic-approximation based adaptive IS techniques for discrete time Markov chains. Applying the adaptive algorithm in Ahamed et al. (2005) to our slotted-time queuing system our main observation is the following. Using the state-dependent change of measure described in Section 4 as an initial condition for stochastic-approximation algorithm leads to very quick convergence of the adaptive algorithm to accurate estimates even for cases where the proposed state-dependent change-of-measure is not effective as stand-alone method for IS.

The algorithm described in Ahamed et al. (2005) adaptively learns the function $J_y(u)$. Using results from Proposition 2, we initialize the proposed algorithm with

$$J_y^{(0)}(u) = \frac{1}{1-\rho}\left[\sum_{x=u-x-y+2}^{u+1} \mathbb{P}(X_1 \geq x)\right].$$

Then, we execute the adaptive algorithm to get an "improved" approximation for the function $J_y(u)$. To construct confidence intervals, we use the approximations of $J_y(u)$ to construct the approximate zero-variance measure for IS as in Section 4. To study, the effectiveness of this approach, we apply this to the parameters values on our numerical example where the proposed state-dependent IS change-of-measure does not perform well; see table 1 where these values are marked with '†'. The number of iterations for the adaptive algorithm is taken to be $200,000$ and the simulation runs for the resulting IS estimator is taken to be $300,000$. (Thus, the computational effort is at par with the earlier experiments.) The results are displayed in table 2. We observe that the aforementioned approach improves the accuracy of the estimator by an order of magnitude.

| $(\alpha, \rho)$ | Probability estimate |
|---|---|
| $(10, 0.8)$ | $7.5156 \times 10^{-19} \pm 7.6\%[1.83]$ |
| $(20, 0.5)$ | $1.5951 \times 10^{-41} \pm 3.2\%[1.93]$ |
| $(20, 0.8)$ | $1.0252 \times 10^{-19} \pm 4.9\%[1.85]$ |

Table 2: IS estimator of buffer overflow probability during a busy cycle: Simulation results obtained using a combination of adaptive algorithm and the state-dependent IS distribution proposed in Section 4. The buffer level is 100 and the number of iteration for the adaptive algorithm is $200,000$ followed by $300,000$ simulation runs. The number in square parenthesis represents the $[Y]$ represents the ratio defined in (3).

## 5 Concluding Remarks

**1.** Theorems 1 and 2 imply that for our class of heavy-tailed distributions no state-independent change-of-measure can be asymptotically optimal, since by Definition 1 such a distribution must satisfy

$$\liminf_{u\to\infty} \frac{\log \widetilde{\mathbb{E}}[Z_u^2]}{-\alpha \log u} \geq 2.$$

Theorems 1 and 2 can be seen to hold even when the IS distribution is allowed to depend on $u$, and Theorem 2 continues to hold when the inter-arrival time distribution is changed in a state-dependent manner.

**2.** The bounds given in Theorems 1 and 2 indicate that the efficiency loss corresponding to the "best" state-independent IS distribution is more severe the heavier the tails of the underlying distributions are. As these tails become lighter, a state-independent IS distribution may potentially achieve near-optimal asymptotic variance reduction.

**3.** As noted earlier, in both cases covered in Theorems 1 and 2 there exists a zero-variance IS distribution that has a "Markovian structure." This suggests that an implementable good approximation to this measure may be feasible and may serve as an effective state-dependent IS distribution. We provided an illustration of this idea in Section 4 through a simple numerical example. Recently, Blanchet & Glynn (2005) proved asymptotic optimality of such a change-of-measure for estimating the probability that all-time-max of a negative drift random walk exceeds a large threshold. They use a refinement of the asymptote given in Theorem 3 of the appendix to develop an approximate zero variance importance sampling measure.

**4.** The results given in Section 4 suggest that when the original asymptotes are not accurate (and when the refinements are not available), one can "learn" them adaptively to devise a good state-dependent IS measure. The extension of the work in Ahamed et al. (2005) to cover general state-space is pursued in separate work.

## A Proofs

**Proof of Proposition 1.** Let $W_i$ be the waiting time of the $i^{th}$ arrival, and let $M_n = \max_{1\leq m\leq n} W_m$. Thus, $M_{N(\tau)}$ is the maximum waiting time during the busy

cycle. Consider the following inequalities

$$\begin{aligned}
&\mathbb{P}\left(\max_{0\le t\le\tau} Q(t) > u\right)\\
&\quad\geq \mathbb{P}\left(\max_{0\le t\le\tau} Q(t) > u,\ M_{N(\tau)} > 2\lambda^{-1}u\right)\\
&\quad= \mathbb{P}\left(M_{N(\tau)} > 2\lambda^{-1}u\right)\times\\
&\qquad\mathbb{P}\left(\max_{0\le t\le\tau} Q(t) > u | M_{N(\tau)} > 2\lambda^{-1}u\right).
\end{aligned}$$

Now consider the second term on right-hand-side and the set of paths where the arrivals during the "large" waiting time causes the buffer to overflow. That is, conditioned on the event that there exists a customer that experience a large waiting time, we consider the arrivals which take place while this customers waits. A sufficient condition for overflow is that the sum of the next $u$ inter-arrival times is less than the waiting time $M_{N(\tau)}$. Since inter-arrival times are iid we have,

$$\begin{aligned}
&\mathbb{P}\left(\max_{0\le t\le\tau} Q(t) > u | M_{N(\tau)} > 2\lambda^{-1}u\right)\\
&\qquad\geq \mathbb{P}\left(\sum_{i=1}^{\lfloor u\rfloor} A_i \le 2\lambda^{-1}u\right),
\end{aligned}$$

where $\{A_i, i = 1, 2, \ldots\}$ is the sequence of inter-arrival time r.v.'s. In Asmussen (1998), exact asymptotes for the probability that a random walk hits a large level before it goes below zero are developed. From the asymptote, it can be seen that

$$\log \mathbb{P}\left(M_{N(\tau)} > 2\lambda u\right) \sim -\alpha \log u.$$

Using the strong law of large numbers, we also have

$$\mathbb{P}\left(\sum_{i=1}^{\lfloor u\rfloor} A_i \le 2\lambda^{-1}u\right) \to 1 \text{ as } u \to \infty.$$

Thus, we have

$$\liminf_{u\to\infty} \frac{\log \mathbb{P}(\max_{0\le t\le\tau} Q(t) > u)}{\log u} \geq -\alpha.$$

Now, consider the following bounds

$$\begin{aligned}
&\mathbb{P}\left(M_{N(\tau)} > 0.5\mu^{-1}u\right) &(9)\\
&\quad\geq \mathbb{P}\left(\max_{0\le t\le\tau} Q(t) > u,\ M_{N(\tau)} > 0.5\mu^{-1}u\right) &(10)\\
&\quad= \mathbb{P}\left(\max_{0\le t\le\tau} Q(t) > u\right)\times &(11)\\
&\qquad\mathbb{P}\left(M_{N(\tau)} > 0.5\mu^{-1}u | \max_{0\le t\le\tau} Q(t) > u\right)\\
&\quad\geq \mathbb{P}\left(\max_{0\le t\le\tau} Q(t) > u\right)\mathbb{P}\left(\sum_{i=1}^{\lfloor u\rfloor} S_i > 0.5\mu^{-1}u\right),
\end{aligned}$$

where $\{S_i, i = 1, 2, \ldots\}$ represent the service times r.v.'s. The last inequality follows from the fact that a sufficient condition for the maximum waiting time to exceed $0.5\mu^{-1}u$ is that the sum of service times for customers in the queue exceeds $0.5\mu^{-1}u$. By the strong law of large numbers, we have

$$\mathbb{P}\left(\sum_{i=1}^{\lfloor u\rfloor} S_i > 0.5\mu^{-1}u\right) \to 1 \text{ as } u \to \infty.$$

Thus, using (5) and the random walk representation of the waiting time, we have

$$\limsup_{u\to\infty} \frac{\log \mathbb{P}\left(\max_{0\le t\le\tau} Q(t) > u\right)}{\log u} \le -\alpha.$$

This completes the proof. ■

**Proof of Proposition 2.** Consider the random walk

$$S_n = S_{n-1} - 1 + X_n,$$

so $S_n$ has a negative drift given by $\mathbb{E}X_1 - 1$, and let $\widetilde{\tau}_0 = \inf\{m : S_m \le 0\}$ and $\widetilde{\tau}_u = \inf\{m : S_m \ge u\}$. Note that $J_y(u) = \mathbb{P}(\widetilde{\tau}_0 > \widetilde{\tau}_u | S_0 = y)$ for $y \in [0, u]$. Let $\bar{J}_y(u) = \mathbb{P}(\widetilde{\tau}_u < \infty | S_0 = y)$ for $y \in [0, u]$. Fix $\beta \in (0, 1)$. Since the random walk decreases by at most one unit at any time, we have $S_{\widetilde{\tau}_0} = 0$. Thus

$$\bar{J}_{\beta u}(u) = J_{\beta u}(u) + [1 - J_{\beta u}(u)]\bar{J}_0(u),$$

and rearranging terms we get

$$J_{\beta u}(u) = \frac{\bar{J}_{\beta u}(u) - \bar{J}_0(u)}{1 - \bar{J}_0(u)}.$$

Also, we have $\bar{J}_{\beta u}(u) = \bar{J}_0(u(1-\beta))$. Next we appeal to the following theorem from Asmussen (1987) which gives an asymptotic for $\bar{J}_0(u)$.

**Theorem 3 (Theorem 9.1, Asmussen (1987))** *Consider a random walk $S_n = \sum_{i=1}^n Y_i$ such that $\nu = \mathbb{E}Y_1 < 0$ and $Y_1$ has a cumulative distribution $F$ which is sub-exponential. Let $M = \max_i S_i$, then*

$$\mathbb{P}(M > x) \sim \frac{1}{|\nu|}\int_x^\infty (1 - F(y))dy.$$

Using the asymptote above and Karamata's Theorem (cf. Embrechts et al. (1997)) we get

$$\frac{\bar{J}_0(u)}{\bar{J}_0(u(1-\beta))} \to (1-\beta)^{\alpha-1} \text{ and } \bar{J}_0(u) \to 0 \text{ as } u \to \infty.$$

Also, we have $\mathbb{E}[N] = 1/|\mathbb{E}[X_1] - 1|$. The result follows using the fact that if $a_u \sim b_u$, $c_u \sim d_u$ and $a_u/c_u \to K \in (0, 1)$ as $u \to \infty$ then $(a_u - c_u) \sim (b_u - d_u)$. This completes the proof. ■

## REFERENCES


Ahamed, I., Borkar, V. S. & Juneja, S. (2005), 'Adaptive importance sampling for markov chains using stochastic approximation'. To appear in *Operations Research*.

Asmussen, S. (1987), *Applied Probability and Queues*, John Wiley and Sons, Chichester New York Brisbane Toronto Singapore.

Asmussen, S. (1998), 'Subexponential asymptotics for stochastic processes: extreme behavior, stationary distributions and first passage probabilities', *The Annals of Applied Probability* **8**, 354–374.

Asmussen, S. & Binswanger, K. (1997), 'Ruin probability simulation for subexponential claims', *ASTIN Bull* **27**, 297–318.

Asmussen, S., Binswanger, K. & Hojgaard, B. (2000), 'Rare events simulation for heavy-tailed distributions', *Bernoulli* **6**, 303–322.

Asmussen, S., Kroese, D. P. & Rubinstein, R. Y. (2004), 'Heavy tails, importance sampling and cross-entropy'. preprint.

Bassamboo, A., Juneja, S. & Zeevi, A. (2005), 'On the Efficiency Loss of State-independent Importance Sampling in the Presence of Heavy Tails', To appear in *OR Letters*.

Blanchet, J. & Glynn, P. W. (2005), Efficient rare event simulation for the maximum of random walk with heavy-tailed increments. Applied probability conference, Ottawa.

Boots, N. & Shahabuddin, P. (2001), 'Simulating tail probabilities in GI/GI/1 queues and insurance risk processes with sub-exponential distributions'. preprint.

Embrechts, P., Klppelberg, C. & Mikosch, T. (1997), *Modelling Extremal Events for Insurance and Finance*, Springer-Verlag, Berlin.

Glasserman, P. (2003), *Monte Carlo Methods in Financial Engineering*, Springer-Verlag.

Heidelberger, P. (1995), 'Fast simulation of rare events in queueing and reliability models', *ACM Trans. Modeling Computer Simulation* **5**, 43–85.

Juneja, S. & Shahabuddin, P. (2002), 'Simulating heavy tailed processes using delayed hazard rate twisting', *ACM TOMACS* **12,2**, 94–118.

Kollman, C., Baggerly, K., Cox, D. & Picard, R. (1999), 'Adaptive importance sampling on discrete markov chains', *The Annals of Applied Probability* **9**, 391–412.

Resnick, S. (1997), 'Heavy tail modelling and teletraffic data', *Annals of Statistics* **25**, 1805–1869.

Sadowsky, J. (1991), 'Large deviations and efficient simulation of excessive backlogs in a GI/G/m queue', *IEEE Transactions on Automatic Control* **36**, 1383–1394.

Siegmund, D. (1985), *Sequential Analysis: Tests and Confidence Intervals*, Springer-Verlag, New York.


## AUTHOR BIOGRAPHIES


**ACHAL BASSAMBOO** is Donald P. Jacobs Scholar in managerial economics and decision sciences at Kellogg School of Management. His research interests are stochastic systems, revenue management and rare event simulation. His e-mail address is `a-bassamboo@northwestern.edu`.

**SANDEEP JUNEJA** is a faculty in the School of Technology and Computer Science at the Tata Institute of Fundamental Research. His research interests include Monte-Carlo methods for stochastic systems. His e-mail address is `juneja@tifr.res.in` and his web page is `www.tcs.tifr.res.in/~sandeepj`.

**ASSAF ZEEVI** is Nathan Gantcher Associate Professor of Business in the Graduate School of Business, Columbia University. His main research focuses on stochastic models of service systems, with other recent research addressing problems in financial economics, simulation, statistics and applied probability. His email address is `assaf@gsb.columbia.edu`

| u | $\alpha$ | $\rho = 0.3$ | $\rho = 0.5$ | $\rho = 0.8$ |
|---|---|---|---|---|
| 100 | 2 | $3.31 \times 10^{-6} \pm 0.019\%$ [1.97] | $2.43 \times 10^{-5} \pm 0.050\%$ [1.86] | $1.00 \times 10^{-4} \pm 0.837\%$ [1.26] |
| | 9 | $1.57 \times 10^{-23} \pm 0.051\%$ [1.97] | $1.52 \times 10^{-20} \pm 0.119\%$ [1.93] | $5.12 \times 10^{-19} \pm 2.409\%$ [1.79]$^{\dagger}$ |
| | 19 | $4.70 \times 10^{-48} \pm 0.080\%$ [1.98] | $5.30 \times 10^{-42} \pm 0.543\%$ [1.94]$^{\dagger}$ | $4.58 \times 10^{-39} \pm 4.182\%$ [1.89]$^{\dagger}$ |
| 1000 | 2 | $3.19 \times 10^{-8} \pm 0.006\%$ [1.99] | $2.25 \times 10^{-7} \pm 0.015\%$ [1.98] | $8.16 \times 10^{-7} \pm 0.079\%$ [1.84] |
| | 9 | $1.02 \times 10^{-32} \pm 0.007\%$ [2.00] | $9.21 \times 10^{-30} \pm 0.032\%$ [1.99] | $2.49 \times 10^{-28} \pm 0.103\%$ [1.96] |
| | 19 | $7.22 \times 10^{-68} \pm 0.022\%$ [2.00] | $6.72 \times 10^{-62} \pm 0.041\%$ [2.00] | $3.30 \times 10^{-59} \pm 0.403\%$ [1.96] |

Table 1: Performance of the state-dependent IS estimator for buffer overflow probability in a busy cycle: Simulation results for buffer levels 100 and 1000, using $500,000$ simulation runs. 95% confidence intervals are provided. The number in the square parenthesis represents the ratio defined in (3). Recall that 2 corresponds to the asymptotically optimal value.

$^{\dagger}$ The actual probability, which is calculated using first step analysis, lies outside the 95% confidence interval.