Rare-event Simulation for Multistage Production-inventory Systems

Paul Glasserman • Tai-Wen Liu
Columbia Business School, New York, New York 10027
Rutgers University, Newark, New Jersey 07102

We consider the problem of precise estimation of service-level measures in multistage production-inventory systems when the system is managed for high levels of service. Precisely because the service level is high, stockouts, large backorders, and unfilled demands are rare and thus difficult to estimate by straightforward simulation. We propose and analyze alternative estimators, based on changing the demand distribution to make these rare events less rare. Whereas straightforward simulation for a fixed relative error results in computational requirements that grow exponentially in certain stock-level parameters, the requirements for our *importance sampling* estimators remain bounded for all parameter values. We provide bounds making it possible to determine the maximum number of replications required before any are generated. Numerical examples illustrate the effectiveness of our method.
(*Capacity-constrained Inventory Models; Multiechelon Inventory Systems; Service-level Estimation; Rare Events; Tail Probabilities*)

1. Introduction

A primary objective of production and inventory management is the attainment of a high level of service. Important indicators of service level include the frequency of stockouts, the average number of unfilled orders, and the proportion of demands not met from stock; an effective inventory policy keeps each of these quantities low without incurring excessive holding hosts. In a complex, multistage production system, the evaluation of service measures and costs often requires simulation. Somewhat ironically, the very effectiveness of an inventory policy can make accurate estimation of service measures difficult by making deviations from ideal service exceedingly rare.

We investigate the problem of precise service-level estimation in a class of multistage production-inventory systems for which simulation is the only available numerical method. We consider multiple facilities in series, each intermediate node drawing raw material from its predecessor and supplying its successor, the lowest node supplying external demands. Production decisions and the movement of inventory are determined by a *base-stock* policy for *echelon* inventory in which each node attempts to restore cumulative downstream inventory to a target level, called the *base-stock level*. The model is rendered analytically intractable primarily by the presence of production capacity limits at each stage. Actual production at each stage may fall short of target production because of either the unavailability of upstream inventory or the production capacity constraints. Clark and Scarf (1960), Rosling (1989), and Langenhoff and Zijm (1992) study similar models, but without capacity constraints. Federgruen and Zipkin (1986) and Tayur (1992) study the single-stage capacitated case. Our simulation analysis starts from recursions introduced in Glasserman and Tayur (1994) and builds on asymptotics and bounds in Glasserman (1993).

We summarize the evolution of the system through a vector *shortfall* process recording the production deficit for each stage. From the stationary shortfall distribution we can obtain various measures of performance—in particular, the stockout probability, the average backlog, and the fill rate, which is the proportion of demands met directly from stock. At high base-stock levels, stock-

0025-1909/96/4209/1292$01.25

outs are rare and all these measures of service become difficult to estimate. Indeed, we show that using straightforward simulation the number of replications required to achieve a fixed accuracy grows exponentially in the lowest base-stock level. We address this problem by introducing *importance sampling* estimators. Our method simulates shortfalls with a new demand distribution, under which stockouts become less rare and all three service measures are efficiently estimated. Our estimators are unbiased and have *bounded relative error*, meaning that the number of replications required to achieve a fixed accuracy is independent of the base-stock levels. We provide bounds that make it possible to determine a priori the maximal number of replications required.

Our use of importance sampling continues a line of development that includes Siegmund (1976), Asmussen (1987, §XII.7, and 1990), Sadowsky (1991), Lehtonen and Nyrhinen (1992), and Chang et al. (1994), among others; see Heidelberger (1995) for an overview of rare-event simulation. These results exploit exponential asymptotics to identify effective importance sampling distributions, as we do. (Different asymptotics corresponding to a different class of rare events are used for the same purpose in Shahabuddin 1994, where the concept of bounded relative error is introduced.) Much of this work deals with level-crossing probabilities for one-dimensional processes. Our application differs in that the shortfall process we study is vector-valued and, except for the stockout probability, the quantities we estimate are not expressible as level-crossing probabilities. These differences result in somewhat different arguments to establish bounded relative error and to identify explicit bounds.

Although we consider only a specific class of models, our approach should be applicable to other production-inventory systems. A relatively minor extension of our analysis could be used to treat models with variable production capacity, fixed leadtimes, and assembly topologies, because the asymptotics in Glasserman (1993) apply to these systems. More generally, our implementation of importance sampling through a particular change in the demand distribution seems likely to be effective for many systems with limited production capacity. It is also worth noting that though our analysis is asymptotic, numerical experiments suggest that our estimators result in substantial improvement in simulation efficiency even at moderate parameter values.

The rest of this paper is organized as follows. Section 2 specifies the model details and gives the shortfall recursions on which our analysis is based. Section 3 explains the problem of rare events and shows that straightforward simulation is inadequate. We present our estimators and state our main results in §4. The analysis of the estimators and all proofs are deferred to §5.

2. The Model

We consider a multistage production-inventory system with d nodes in series. Time is divided into periods of fixed length, with inventory levels reviewed and production decisions made once in each period. Node i, $i = 1, \ldots, d-1$, draws material from node $i+1$, and node d draws from an unlimited source of raw material. Each node has limited production capacity. Let c^i denote the capacity at node i, and I_n^i the net inventory at stage i at the end of period n. Demands for finished goods arrive in each period and are either filled or backlogged, depending on the net inventory I_n^1 at stage 1. After the total demand D_n in period n is revealed, node i sets its production level to try to restore the *cumulative* inventory $\Sigma_{j=1}^i I_n^j$ to a specified level s^i, called the base-stock level for *echelon* i. (Echelon i refers to the subsystem consisting of stages 1 through i.) In short, node i follows a base-stock policy for echelon inventory with base-stock level s^i for all i. However, two constraints potentially limit production: either the production capacity or the unavailability of upstream inventory can prohibit a full restoration of target inventory in a single period. Closely related variants of this model include systems with imperfect production, fixed leadtimes between stages, and an assembly system with more general topologies. See Glasserman and Tayur (1994) and Glasserman (1993) for discussions of these variants.

Define the shortfall Y_n^i for echelon i to be the amount by which the net echelon inventory falls short of the target level s^i at the end of period n; that is, $Y_n^i = s^i - \Sigma_{j=1}^i I_n^j$. Under a base-stock policy, stage i sets production to drive Y_n^i to zero, while not exceeding

its production capacity c^i or the available upstream inventory I_{n-1}^{i+1}.

The dynamics of the system can be fully captured by the shortfall vector process $Y = \{(Y_n^1, \ldots, Y_n^d), n \geq 0\}$. Glasserman and Tayur (1994) show that the shortfalls satisfy

$$Y_n^d = \max\{0, Y_{n-1}^d + D_n - c^d\}; \tag{1}$$

$$Y_n^i = \max\{0, Y_{n-1}^i + D_n - c^i, Y_{n-1}^{i+1} + D_n - (s^{i+1} - s^i)\}, \tag{2}$$

$i = 1, \ldots, d-1$, and that the vector process $\{Y_n, n \geq 1\}$ admits a finite stationary distribution to which it converges from all initial distributions, provided that the demands $\{D_n, n \geq 1\}$ are (nonnegative and) i.i.d. with distribution F_D and

$$\mathsf{E}[D_1] < c^* \equiv \min\{c^i: 1 \leq i \leq d\}. \tag{3}$$

This result underlies our use of the regenerative method in §4.

As shown in Glasserman (1993), various key measures of performance can be represented in terms of Y^1, a random variable with the stationary distribution of $\{Y_n^1, n \geq 0\}$. The *stockout probability*, or the long-run average proportion of periods in which stockout occurs, is

$$\alpha(s^1) = P\{Y^1 > s^1\}.$$

The long-run expected *average backlog* is

$$b(s^1) = \mathsf{E}(Y^1 - s^1)^+,$$

where $x^+ \equiv \max\{0, x\}$. The *fill rate*, or the long-run average proportion of demands met from stock, can be seen to be

$$\beta(s^1) = 1 - \frac{\mathsf{E}[\{\min(Y^1 + D - s^1, D)\}^+]}{\mathsf{E}[D]}.$$

Though we focus exclusively on these *service-level* measures, it is worth noting that our analysis is relevant to cost measures as well. If, for example, echelon-i inventory is charged a holding cost h_i and if backorders are penalized at rate p, then the long-run average cost per period is

$$\sum_{i=1}^{d} h_i(s^i - \mathsf{E}[Y^i]) + \left(p + \sum_{i=1}^{d} h_i\right)\mathsf{E}[(Y^1 - s^1)^+],$$

so evaluation of $b(s^1)$ is essential to the evaluation of the average cost. (Evaluation of $\mathsf{E}[Y^i]$ does not pose a particular problem for simulation.)

3. The Problem of Rare Events

As the base-stock level s^1 becomes large, stockouts become infrequent, the average backlog becomes almost negligible, and in the long run most demands can be filled from stock. Thus, with a high base-stock level s^1, each of these measures is related to a rare event. (Only $\alpha(s^1)$ is the probability of an event, but we use the term *rare event* loosely to refer to quantities that vanish as s^1 increases.) The degree of rarity is characterized by asymptotics and bounds in Glasserman (1993) for single- and multistage systems. Under conditions to be spelled out shortly, Theorem 1 of Glasserman asserts that for a single-stage system with capacity c and target level s, there are constants $\gamma > 0$ and $C > 0$ such that

- the stockout probability satisfies $\alpha(s) \sim Ce^{-\gamma s}$;
- the average backlog satisfies $b(s) \sim (C/\gamma)e^{-\gamma s}$;
- the fill rate satisfies $1 - \beta(s) \sim (C/\gamma\mathsf{E}[D])(e^{\gamma c} - 1)e^{-\gamma s}$.

The notation $f(x) \sim g(x)$ means that $f(x)/g(x)$ converges to unity as $x \to \infty$.

To see the implications of these asymptotics for simulation, consider the case of $\alpha(s)$. The relation $\alpha(s) \sim Ce^{-\gamma s}$ means that $\alpha(s)$ drops off exponentially at rate γ as the base-stock level s increases, making stockouts exceedingly rare for large s. The most straightforward method of estimating $\alpha(s)$ generates independent realizations of the indicator $\mathbf{1}_{\{Y>s\}}$ and averages them (assuming, for simplicity, that Y can be sampled directly). Let $\bar{\alpha}_n(s)$ be the sample mean of n such realizations. Evidently, $\mathsf{E}[\bar{\alpha}_n(s)] = \alpha(s)$ and

$$\mathbf{Var}[\bar{\alpha}_n(s)] = n^{-1}[\alpha(s) - (\alpha(s))^2] \approx n^{-1}\alpha(s),$$

because $\alpha(s)$ is small. Define the relative error, RE, of an estimator to be the ratio of its standard error to its mean. Then

$$\mathrm{RE}(\bar{\alpha}_n(s)) = \frac{\sqrt{\mathbf{Var}[\bar{\alpha}_n(s)]}}{\alpha(s)} \approx \sqrt{\frac{\exp\{\gamma s\}}{nC}},$$

according to the asymptotic approximation for $\alpha(s)$ given above. For fixed n, as the event becomes rarer (i.e., $s \to \infty$) the RE becomes unbounded. For fixed s, suppose we want to choose n to achieve a relative error of δ, $0 < \delta < 1$; then, by setting $\sqrt{\exp(\gamma s)/nC} = \delta$, we find that we must have roughly $\exp(\gamma s)/(\delta^2 C)$ replications of $\mathbf{1}_{\{Y>s\}}$. Thus, as the target level s increases, the number of runs required grows exponentially, rendering the method infeasible. This is the major problem with straightforward simulation.

We address this problem by using importance sampling to develop estimators for which the RE is bounded uniformly in s. The number of runs required to achieve a specified accuracy thus remains bounded as s increases. Our estimators are based on simulating the system with a new demand distribution under which stockouts become less rare. To make this precise, we introduce some additional assumptions on the original demands. By supposing that $P\{D_1 > c^*\} > 0$ we exclude the trivial case where demands may always be met from the current period's production. For simplicity we also assume that $D_1 - c^*$ is nonarithmetic; otherwise, the asymptotics given above hold only through appropriate subsequences. Our most important assumption is that there exists a positive θ_0 at which

$$1 < \phi(\theta_0) \equiv \mathsf{E}[e^{\theta_0(D_1 - c^*)}] < \infty. \tag{4}$$

This condition, together with the convexity of the moment generating function ϕ, and the properties $\phi(0) = 1$ and $\phi'(0) = \mathsf{E}[D_1 - c^*] < 0$, guarantees the existence of a unique $\gamma > 0$, called the *conjugate point* for the distribution of $D_1 - c^*$, which solves

$$\phi(\gamma) = 1. \tag{5}$$

From F_D and γ we define a new distribution $\tilde{F}_D$ by setting

$$\tilde{F}_D(t) = \int_0^t e^{\gamma(u - c^*)} dF_D(u), \qquad t > 0.$$

Condition (5) ensures that $\tilde{F}_D$ is indeed a probability distribution. Condition (4) is satisfied by most standard distributions with exponential tails.

Our estimators rely on the distribution $\tilde{F}_D$ and an association between shortfall processes and random walks. Specifically, let $X_n = D_n - c^*$, and

$$S_n = X_1 + \cdots + X_n, \; n \geq 1, \quad \text{with } S_0 = 0. \tag{6}$$

Because $\mathsf{E}[X_1] = \mathsf{E}[D_1] - c^* < 0$, the random walk S_n has negative drift. If we let $\{\tilde{D}_n, n \geq 1\}$ have distribution $\tilde{F}_D$, and set $\tilde{X}_n = \tilde{D}_n - c^*$, then $\{\tilde{X}_n, n \geq 1\}$ is a sequence of i.i.d. random variables with distribution $\tilde{F}_X(x) = \tilde{F}_D(x + c^*)$, and

$$\tilde{S}_n = \tilde{X}_1 + \cdots + \tilde{X}_n, \; n \geq 1, \quad \text{with } \tilde{S}_0 = 0, \tag{7}$$

is the conjugate random walk associated with $\{S_n, n \geq 0\}$. The conjugate walk $\tilde{S}_n$ has positive drift because $\mathsf{E}[\tilde{X}_1] = \phi'(\gamma) > 0$. (Condition (4) ensures that $\tilde{X}_1$ has finite moments of all order; that is, $\mathsf{E}[\tilde{X}_1^k] = \mathsf{E}[X_1^k e^{\gamma X_1}] = \phi^{(k)}(\gamma) < \infty$ for every positive integer k.) Thus, demands drawn from $\tilde{F}_D$ have mean greater than c^* and lead to frequent stockouts.

4. Main Results

We now present estimators with bounded relative error for the performance measures defined in §2. Each of our algorithms simulates shortfalls with the new demand distribution $\tilde{F}_D$ until some stopping time τ, then either stops or continues with the original demand distribution F_D. The choice of stopping time depends on the problem at hand. In an effort to unify notation we introduce three auxiliary systems of recursions. First, we let the vector process $(\tilde{Y}_n^1, \ldots, \tilde{Y}_n^d)$ with $(\tilde{Y}_0^1, \ldots, \tilde{Y}_0^d) = (0, \ldots, 0)$ mirror the form of (1)–(2) with D_n replaced by $\tilde{D}_n$ up to some stopping time τ; that is,

$$\tilde{Y}_n^d = \begin{cases} \max\{0, \tilde{Y}_{n-1}^d + \tilde{D}_n - c^d\} & \text{for } 1 \leq n \leq \tau, \\ \max\{0, \tilde{Y}_{n-1}^d + D_n - c^d\} & \text{for } n > \tau, \end{cases} \tag{8}$$

and

$$\tilde{Y}_n^i = \begin{cases} \max\{0, \tilde{Y}_{n-1}^i + \tilde{D}_n - c^i, \tilde{Y}_{n-1}^{i+1} \\ \qquad + \tilde{D}_n - (s^{i+1} - s^i)\} & \text{for } 1 \leq n \leq \tau, \\ \max\{0, \tilde{Y}_{n-1}^i + D_n - c^i, \tilde{Y}_{n-1}^{i+1} \\ \qquad + D_n - (s^{i+1} - s^i)\} & \text{for } n > \tau, \end{cases} \tag{9}$$

for $i = 1, \ldots, d-1$.

The second and the third sets of recursions correspond to unreflected shortfalls, meaning that the zero components in the recursions are left out. Let

$$(S_0^1, \ldots, S_0^d) = (\tilde{S}_0^1, \ldots, \tilde{S}_0^d) = (0, \ldots, 0).$$

The vector $(S_n^1, \ldots, S_n^d)$ satisfies the following:

$$S_n^d = S_{n-1}^d + D_n - c^d; \tag{10}$$

$$S_n^i = \max\{S_{n-1}^i + D_n - c^i, S_{n-1}^{i+1} + D_n - (s^{i+1} - s^i)\}, \tag{11}$$

for $i = 1, \ldots, d-1$ and all $n \geq 1$. Similarly, we set

$$\tilde{S}_n^d = \tilde{S}_{n-1}^d + \tilde{D}_n - c^d; \tag{12}$$

$$\tilde{S}_n^i = \max\{\tilde{S}_{n-1}^i + \tilde{D}_n - c^i, \tilde{S}_{n-1}^{i+1} + \tilde{D}_n - (s^{i+1} - s^i)\}, \tag{13}$$

for $i = 1, \ldots, d-1$. The one-dimensional random walk $\{S_n\}$ (respectively, $\{\tilde{S}_n\}$) is not to be confused with the vector process $(S_n^1, \ldots, S_n^d)$ (respectively, $(\tilde{S}_n^1, \ldots, \tilde{S}_n^d)$), though S_n coincides with S_n^d (respectively, $\tilde{S}_n$ with $\tilde{S}_n^d$) in the important special case that $c^* = c^d$. All the recursions (8)–(13) are easily implemented in a simulation algorithm. (The question of how best to sample from $\tilde{F}_D$ depends, of course, on the original distributions F_D. Acceptance-rejection schemes for some particular cases are analyzed in Nakayama 1992.)

4.1. Stockout Probability

Because of a nice link to a stopping time, the stockout probability $\alpha(s^1) = P\{Y^1 > s^1\}$ admits a straightforward simulation procedure involving the unreflected shortfalls $(\tilde{S}_n^1, \ldots, \tilde{S}_n^d)$ in (12)–(13) and the random walk $\tilde{S}_n$ in (7). Specifically, from recursions (1)–(2) and (10)–(11), it follows that $Y^1 =_d \max_{n \geq 0} S_n^1$, where $=_d$ denotes equality in distribution. Let

$$\tilde{T}(s^1) = \inf\{n \geq 1: \tilde{S}_n^1 > s^1\}, \tag{14}$$

and define $T(s^1)$ analogously for S_n^1. Then

$$P\{Y^1 > s^1\} = P\{T(s^1) < \infty\} = \mathsf{E}[e^{-\gamma \tilde{S}_{\tilde{T}(s^1)}}], \tag{15}$$

the last equality following from Wald's likelihood ratio identity (as in Asmussen 1987, p. 266). (See the proof of Theorem 1 for details.) An unbiased estimator now emerges:

Stockout Estimation.

1. Simulate $(\tilde{S}_n^1, \ldots, \tilde{S}_n^d)$ until $\tilde{T}(s^1)$.
2. Return the estimator $\exp\{-\gamma \tilde{S}_{\tilde{T}(s^1)}\}$.

In this algorithm the demands are always generated from $\tilde{F}_D$. It is worth noting that the resulting estimator differs in an essential way from related estimators for one-dimensional random walks (such as the one in Siegmund 1976): the process $\tilde{S}^1$ defining the level-crossing event in (14) is not a random walk, nor does it coincide with the process $\tilde{S}$ appearing in the estimator. Nevertheless, we have

THEOREM 1. *The estimator* $\exp\{-\gamma \tilde{S}_{\tilde{T}(s^1)}\}$ *is unbiased and has bounded RE. The RE is bounded above by* $\sqrt{C_+} e^{\gamma(\zeta_+ - \zeta_-)}/C_-$, *for constants* C_-, C_+, ζ_- *and* ζ_+ *given in* (30)–(33).

REMARK 1. (i) The availability of an explicit upper bound on the relative error is useful in simulation planning: if the RE for one replication is bounded by a constant A, then the number of replications required to achieve a relative error of δ is bounded by $(A/\delta)^2$, and this can be determined before any replications are generated. Of course, the efficiency of the estimator does not rely on the availability of an explicit upper bound. (ii) In certain exceptional cases, the constant C_- may be zero, rendering the bound in the theorem vacuous. Even in this case, a finite bound is available. We postpone the details of this case to Remark 4 of §5.

4.2. Average Backlog

We give two estimators for the average backlog. The first uses a randomization procedure; the second uses a regenerative representation.

By recalling the definition of the average backlog, we find that it can be written as

$$b(s^1) \equiv \mathsf{E}(Y^1 - s^1)^+ = \int_{s^1}^{\infty} P\{Y^1 > x\} dx.$$

The rightmost expression, combined with the estimator for stockout probability given earlier, suggests a first estimator. To obtain an implementable procedure, we replace the infinite upper limit of integration with a random upper limit L, thus adapting a method introduced by Asmussen (1990) for heavy-traffic simulation. The algorithm is as follows.

Backlog Estimation: Randomization Method.

1. Generate L from an exponential distribution with rate γ.
2. Simulate $(\tilde{S}_n^1, \ldots, \tilde{S}_n^d)$ independently of L until $\tilde{T}(s^1 + L)$, where

$$\tilde{T}(s^1 + L) = \inf\{n \geq 1: \tilde{S}_n^1 > s^1 + L\}. \quad (16)$$

3. Return the estimator

$$e^{-\gamma s^1} \int_{s^1}^{s^1+L} e^{-\gamma(\tilde{S}_{\tilde{T}(x)} - x)} dx. \quad (17)$$

The integral in (17) can be written as a sum for purposes of implementation. To make this explicit, we define

$$t_k = \inf\{n > t_{k-1}: \tilde{S}_n^1 > \tilde{S}_{t_{k-1}}^1\}, \qquad k = 1, \ldots, m.$$

Set $\tilde{T}(s^1) = t_0$ and $\tilde{T}(s^1 + L) = t_m$. Then

$$\tilde{S}_{\tilde{T}(x)} - x = \begin{cases} \tilde{S}_{t_0} - x & \text{for } s^1 \leq x < \tilde{S}_{t_0}^1, \\ \tilde{S}_{t_k} - x & \text{for } \tilde{S}_{t_{k-1}}^1 \leq x < \tilde{S}_{t_k}^1, \\ \quad k = 1, \ldots, m-1, & \\ \tilde{S}_{t_m} - x & \text{for } \tilde{S}_{t_{m-1}}^1 \leq x < s^1 + L. \end{cases}$$

By substituting these back into the estimator, it then follows that (17) admits the simpler expression

$$\frac{e^{-\gamma s^1}}{\gamma} \left\{ e^{-\gamma \tilde{S}_{t_0}}(e^{\gamma \tilde{S}_{t_0}^1} - e^{\gamma s^1}) + \sum_{k=1}^{m} e^{-\gamma \tilde{S}_{t_k}}(e^{\gamma \tilde{S}_{t_k}^1} - e^{\gamma \tilde{S}_{t_{k-1}}^1}) \right.$$
$$\left. + e^{-\gamma \tilde{S}_{t_m}}(e^{\gamma(s^1+L)} - e^{\gamma \tilde{S}_{t_m}^1}) \right\}.$$

For estimator (17) we have the following result.

THEOREM 2. *The estimator* (17) *is unbiased and has bounded RE. The RE is bounded above by* $\sqrt{\gamma C_+} e^{\gamma(\zeta_+ - \zeta_-)}/C_-$, *for constants* C_-, C_+, ζ_- *and* ζ_+ *given in* (30)–(33).

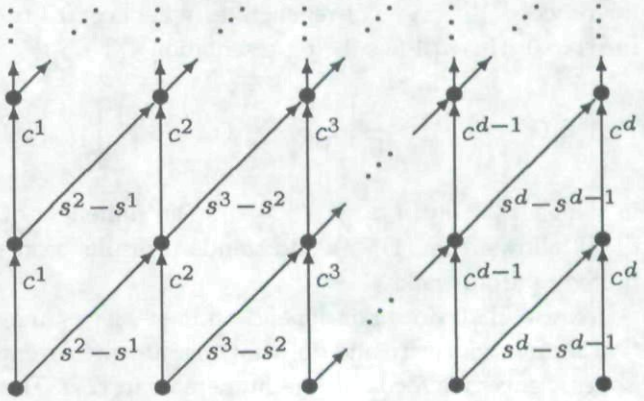
In a related setting, Asmussen (1990) recommends using the random horizon L as a control variate. We examine this option numerically in §4.4.

Our next estimator uses a regenerative framework, so we proceed by identifying regeneration points. This requires a closer examination of the shortfall recursions.

For each k, $1 \leq k \leq d$, let the sequence $\{r_n^k, n \geq 0\}$ denote the length of the shortest n-step path through Figure 1, starting from the lowest node in column k; e.g.,

$$r_2^2 = \min\{2c^2, c^2 + (s^3 - s^2), (s^3 - s^2) + c^3, (s^3 - s^2) + (s^4 - s^3)\},$$

Figure 1 Each Vertical Arc in Column i has Length c^i; Each Diagonal Arc from Column i to Column $i+1$ has Length $s^{i+1} - s^i$



and $r_n^d = nc^d$ for all $n \geq 0$. For convenience we write $r_n = r_n^1$ for all n. For large n, the sequence $\{r_n, n \geq 0\}$ eventually follows a pattern, in spite of an initial irregularity. Indeed, if stages $i_1, i_2, \ldots, i_m$, $(i_1 < \cdots < i_m)$ all have capacity c^*, Glasserman (1993) argues that $r_n - nc^* = \eta$ for all $n \geq n^*$, where

$$\eta \equiv \min_{j \geq i_1}[(s^j - s^1) - (j-1)c^*], \quad (18)$$

and n^* are constants. This means that, for sufficiently large n, the length r_n of the shortest n-step path differs from nc^* only by a constant, which depends on the minimal capacity and the base-stock levels. We use these observations in the following result. We state it for $(Y_n^1, \ldots, Y_n^d)$, but it applies as well to $(\tilde{Y}_n^1, \ldots, \tilde{Y}_n^d)$ defined through (8)–(9).

LEMMA 1. (i) *For* $k = 1, \ldots, d-1$, *we have for all* n

$$Y_n^k + \min_{0 \leq j \leq n}[r_j^k - jc^d] \leq Y_n^d \leq Y_n^k + \max_{0 \leq j \leq n}[r_j^k - jc^d].$$

(ii) *In particular, if*

$$c^d < \min_{i \neq d}\{c^i\} \quad \textit{and} \quad (19)$$

$$c^* \leq s^2 - s^1, \ldots, s^d - s^{d-1}, \quad (20)$$

then $Y_n^k \leq Y_n^d$, *for all* n *and all* k.

While the vector process $\{Y_n, n \geq 0\}$ is a regenerative process whenever the demands satisfy (3), we nonetheless require (19) and (20) to help identify regeneration

points. Under these conditions, it follows from Lemma 1(ii) and the nonnegativity of the shortfalls that the vector process $(Y_n^1, \ldots, Y_n^d)$ regenerates whenever Y^d returns to 0. This justifies the representation

$$\mathsf{E}[(Y^1 - s^1)^+] = \frac{1}{\mathsf{E}[\tau_0^d]}\,\mathsf{E}\left[\sum_{n=0}^{\tau_0^d-1}(Y_n^1 - s^1)^+\right], \quad (21)$$

in which $\tau_0^d = \inf\{n \geq 1\colon Y_n^d = 0\}$. The finiteness of $\mathsf{E}[\tau_0^d]$ follows from (1), (3), and standard results on reflected random walks.

Because $\mathsf{E}[\tau_0^d]$ does not depend on the s^i, its estimation is straightforward and does not raise any rare-event issues; hence, we focus on the numerator in (21). The numerator is nonzero only if Y_n^1 exceeds s^1 in a cycle—a rare event. To make this level-crossing less rare, we first simulate with the new demand distribution $\tilde{F}_D$, thus driving the shortfalls upward. Once s^1 is reached, we suspend the importance sampling to facilitate completion of the cycle. Goyal et al. (1992) call this type of approach *dynamic* importance sampling. The simulation can be carried out as follows.

Backlog Estimation: Regenerative Method

1. Simulate $(\tilde{Y}_n^1, \ldots, \tilde{Y}_n^d)$ until $T_1 \equiv \min\{\tilde{\tau}_0^d, \tilde{\tau}(s^1)\}$, where

$$\tilde{\tau}_0^d = \inf\{n \geq 1\colon \tilde{Y}_n^d = 0\}, \quad \text{and} \quad (22)$$

$$\tilde{\tau}(s^1) = \inf\{n \geq 1\colon \tilde{Y}_n^1 > s^1\}. \quad (23)$$

2. If $\tilde{\tau}(s^1) < \tilde{\tau}_0^d$, continue simulating $(\tilde{Y}_n^1, \ldots, \tilde{Y}_n^d)$, but now using F_D, until $\tilde{\tau}_0^d$.

3. Return the estimator

$$e^{-\gamma \tilde{S}_{\tilde{\tau}(s^1)}} \sum_{n=\tilde{\tau}(s^1)}^{\tilde{\tau}_0^d-1} (\tilde{Y}_n^1 - s^1)^+. \quad (24)$$

In this setting the stopping time T_1 plays the role played by τ in recursions (8)–(9). The algorithm prescribes that in case $\tilde{\tau}(s^1) < \tilde{\tau}_0^d$, meaning that the process $\tilde{Y}^1$ attains level s^1 in a cycle, we switch from $\tilde{F}_D$ to F_D when the level crossing occurs, simulate until $\tilde{\tau}_0^d$, and then evaluate (24). Otherwise, we must have $\tilde{\tau}_0^d < \tilde{\tau}(s^1)$ resulting in an estimator value of zero. In §5 we verify that $\tilde{\tau}_0^d$ is almost surely finite.

Paralleling earlier results, we have

THEOREM 3. *Estimator* (24) *is unbiased and has bounded RE under conditions* (19)–(20). *The RE is bounded above by* $\gamma\sqrt{B}e^{\gamma\zeta_+}/C_-$, *for constants* C_-, ζ_+ *and* B *given in* (30), (33), *and* (46).

A regenerative importance sampling estimator for stockout probability could be defined through a minor modification of (24). However, such an estimator seems inferior to the one defined in §4.1. On any cycle in which a stockout occurs, the estimator of §4.1 terminates a replication before the regenerative cycle would be completed. So, there is no incentive to simulate the rest of the cycle.

4.3. Fill Rate

As for the average backlog, we present two estimators for the fill rate. One exploits an integral representation; the other uses the regenerative framework.

First, we use the definition of the fill rate to express the expected demands not filled in a period as

$$\begin{aligned}
&(\mathsf{E}[D])(1 - \beta(s^1)) \\
&\quad = \mathsf{E}[(Y^1 + D - s^1)^+ \mathbf{1}_{\{Y^1 \leq s^1\}}] + \mathsf{E}[D\mathbf{1}_{\{Y^1 > s^1\}}] \\
&\quad = \mathsf{E}[(Y^1 + D - s^1)^+] - \mathsf{E}[(Y^1 - s^1)^+] \\
&\quad = \int_{s^1}^{\infty} (P\{Y^1 + D > x\} - P\{Y^1 > x\})dx. \quad (25)
\end{aligned}$$

From the recursions (1)–(2) and (10)–(11) it follows that

$$Y_{n-1}^1 + D_n =_d \max_{1 \leq j \leq n}[S_{j-1}^1 + D_j],$$

and therefore

$$Y^1 + D =_d \max_{n \geq 1}[S_{n-1}^1 + D_n].$$

By defining $T'(x) = \inf\{n \geq 1\colon S_{n-1}^1 + D_n > x\}$ and $\tilde{T}'(x) = \inf\{n \geq 1\colon \tilde{S}_{n-1}^1 + \tilde{D}_n > x\}$, we have

$$P\{Y^1 + D > x\} = P\{T'(x) < \infty\} = \mathsf{E}[e^{-\gamma \tilde{S}_{\tilde{T}'(x)}}]. \quad (26)$$

The last equality follows from Wald's likelihood ratio identity and the finiteness of $\tilde{T}'(x)$. A combination of (25), (26), and (15) leads to the following algorithm for estimating $(\mathsf{E}[D])(1 - \beta(s^1))$.

Fill-rate Estimation: Randomization Method

1. Generate L from an exponential distribution with rate γ.
2. Simulate $(\tilde{S}_n^1, \ldots, \tilde{S}_n^d)$ independently of L until $\tilde{T}(s^1 + L)$.
3. Return the estimator

$$e^{-\gamma s^1} \int_{s^1}^{s^1+L} [e^{-\gamma(\tilde{S}_{\tilde{T}'(x)} - x)} - e^{-\gamma(\tilde{S}_{\tilde{T}(x)} - x)}]dx. \qquad (27)$$

An argument parallel to the one used for (17) expresses the integral in (27) as a sum.

THEOREM 4. *The estimator* (27) *is unbiased and has bounded RE. If* (20) *holds, the RE is bounded above by*

$$\sqrt{\gamma C_+(e^{2\gamma c^*} + 1)} e^{\gamma(\zeta_+ - \zeta_-)/(1 - C_+)},$$

for constants C_+, ζ_- *and* ζ_+ *given in* (31)–(33).

REMARK 2. (i) Without Condition (20) an upper bound is still available, as explained in Remark 5 of §5. (ii) For a single-stage system with capacity c and base-stock level s, the shortfall recursion $Y_n = \max\{0, Y_{n-1} + D_n - c\}$ (cf. (1)) implies that the steady-state shortfall Y satisfies (see Glasserman 1993)

$$Y =_d \max\{0, Y + D - c\}. \qquad (28)$$

This enables us to write the fill rate as

$$\beta(s) = 1 - \frac{1}{\mathsf{E}[D]} \int_{s-c}^{s} P\{Y > x\}dx,$$

which in turn suggests a simplified estimator for the fill rate in a single-stage case:

$$1 - \frac{1}{\mathsf{E}[D]} \int_{s-c}^{s} e^{-\gamma \tilde{S}_{\tilde{T}_x}} dx,$$

where $\tilde{T}_x$ is the first time the conjugate random walk $\tilde{S}_n$ exceeds level x. It is easy to verify that this estimator is unbiased and has bounded RE.

For our second approach we require assumptions (19)–(20) in order to use a regenerative framework as we did for the average backlog. Let the notation $a \wedge b$ stand for $\min(a, b)$. The expected demands not filled in a period can then be written

$$\begin{aligned}&\mathsf{E}[D](1 - \beta(s^1)) \\ &\quad = \mathsf{E}[(Y^1 + D - s^1)^+ \wedge D] \\ &\quad = \frac{1}{\mathsf{E}[\tau_0^d]} \mathsf{E}\left\{ \sum_{n=1}^{\tau_0^d - 1} [(Y_{n-1}^1 + D_n - s^1)^+ \wedge D_n] \right\},\end{aligned}$$

where τ_0^d is again the first time the vector process returns to the origin. We use dynamic importance sampling to simulate the numerator as follows:

Fill-rate Estimation: Regenerative Method

1. Simulate $(\tilde{Y}_n^1, \ldots, \tilde{Y}_n^d)$ until $T_2 = \min\{\tilde{\tau}_0^d, \tilde{\tau}'(s^1)\}$, with $\tilde{\tau}_0^d$ as in (22), and

$$\tilde{\tau}'(s^1) = \inf\{n \geq 1: \tilde{Y}_{n-1}^1 + \tilde{D}_n > s^1\}.$$

2. If $\tilde{\tau}'(s^1) < \tilde{\tau}_0^d$, continue simulating $(\tilde{Y}_n^1, \ldots, \tilde{Y}_n^d)$, but now using F_D, until $\tilde{\tau}_0^d$.
3. Return the estimator

$$e^{-\gamma \tilde{S}_{\tilde{T}'(s^1)}} \sum_{n=\tilde{T}'(s^1)}^{\tilde{\tau}_0^d - 1} [(\tilde{Y}_{n-1}^1 + D_n - s^1)^+ \wedge D_n], \qquad (29)$$

where $\tilde{Y}_n^d$ and $\tilde{Y}_n^1$ satisfy the recursions (8)–(9) with $\tau = T_2$.

THEOREM 5. *The estimator* (29) *is unbiased and has bounded RE under conditions* (19)–(20). *The RE is bounded above by* $\gamma\sqrt{B} e^{\gamma(\zeta_+ + c^d)}/(1 - C_+')$, *for constants* ζ_+, C_+, *and* B *given in* (33), (31), *and* (46).

REMARK 3. In light of (28) the fill rate in a single-stage system has the simpler representation

$$\beta(s) = 1 - \frac{1}{\mathsf{E}[D]} \mathsf{E}[(Y - (s - c))^+ \wedge c],$$

from which a simplified regenerative estimator follows.

4.4. Numerical Results

We have experimented at several parameter settings with the estimators of §§4.1–4.2 for stockout probability and average backlog. In the case of average backlog, we can compare the performance of the randomization and regenerative estimators; in all cases we compare the performance of our importance sampling estimators with a standard estimate of steady-state performance based on regenerative cycles. All experiments were carried out on a 386 PC using the SIMAN simulation language. Tables 1 and 2 summarize the estimates obtained from the

Table 1 **Comparison of Importance Sampling and Standard Estimates for Stockout Probability in a Two-stage System with Exponential Demands. The Execution Time is about 14 Minutes for $\rho = 0.60$, 17 Minutes for $\rho = 0.8$, and 25 Minutes for $\rho = 0.98$**

			IS		Standard	
ρ	s^1	Exact	Est.	S.D.	Est.	S.D.
0.60	1	0.01561	0.01663	6.42E−4	0.00848	4.16E−3
	3	0.00132	0.00135	4.53E−5	N/A	N/A
	5	0.000128	0.000127	3.87E−6	N/A	N/A
0.80	1	0.1649	0.1662	2.20E−3	0.1901	3.76E−2
	3	0.0624	0.0622	7.95E−4	0.0810	2.81E−2
	7	0.00964	0.00955	1.26E−4	0.01274	1.01E−2
0.98	30	0.2623	0.2599	2.76E−3	0.2191	9.90E−2
	45	0.14276	0.13885	2.59E−3	0.13605	9.07E−2
	60	0.0777	0.07564	1.62E−3	0.07052	5.90E−2

different methods using a fixed computing time for each estimator. The model under consideration is a two-stage system with $c^1 = 2$, $c^2 = c^* = 1$, $\Delta = s^2 - s^1 = 3$ and exponential demands. For this model, the exact value of both measures, dependent on s^1, could be computed analytically. The tables show results for three levels of the utilization parameter $\rho \triangleq \mathrm{E}[D]/c^*$.

In Table 1, our importance sampling (IS) estimator for stockout probability is contrasted to the standard regenerative estimator. Each row shows results for the value of s^1 specified in the second column. The exact value is given in the third column. Each cell records a point estimate and estimated standard error. The notation "N/A" indicates that all regenerative cycles yielded a value of zero. The results in Table 1 indicate substantial variance reduction from importance sampling. The impact of bounded relative error is most notable at $\rho = 0.6$, where each order-of-magnitude decrease in the stockout probability is accompanied by an order-of-magnitude decrease in the standard deviation, keeping the ratio roughly constant.

Table 2 reports results for average backlog. Four estimators are compared: the regenerative importance sampling estimator (IS/Reg), the randomized importance sampling estimator (IS/Rand), the randomized estimator with L as a control variate (IS/Rand/CV), and the standard regenerative method. Though it might be argued that a control variate could be used with any of the methods, it seems most natural to examine its effectiveness in the randomized method. Randomization introduces new variability into the simulation; using L as a control removes this extra variability.

Table 2 again shows the potential for significant variance reduction via importance sampling, with the possible exception of the IS/Reg estimate at the extreme utilization of $\rho = 0.98$. When backorders are rare (the primary focus of this study), the three importance sam-

Table 2 **Comparison of Three Importance Sampling Estimates and a Standard Estimate for Average Backlog in a Two-stage System with Exponential Demands. The Execution Time is About 16 Minutes for $\rho = 0.60$, 24 Minutes for $\rho = 0.8$, and 30 Minutes for $\rho = 0.98$**

			IS/Reg		IS/Rand		IS/Rand/CV		Standard	
ρ	s^1	Exact	Est.	S.D.	Est.	S.D.	Est.	S.D.	Est.	S.D.
0.60	1	0.0125	0.0126	1.82E−3	0.0125	5.34E−4	0.0127	4.01E−4	0.0083	4.99E−3
	3	0.0011	0.0010	8.88E−5	0.0012	5.11E−5	0.0012	3.41E−5	N/A	N/A
	5	0.000112	0.000083	7.19E−6	0.000105	4.35E−6	0.00011	2.93E−6	N/A	N/A
0.80	1	0.3434	0.3609	5.43E−2	0.3580	1.15E−2	0.3490	2.99E−3	0.3568	1.11E−1
	3	0.1335	0.1474	2.19E−2	0.1339	4.45E−3	0.1338	1.30E−3	0.1409	7.10E−2
	7	0.02076	0.01651	2.30E−3	0.02109	6.97E−4	0.02052	2.1E−4	0.02349	2.20E−2
0.98	30	6.4680	5.5657	5.5348	6.0654	1.0018	6.4660	0.00999	5.9622	3.8870
	45	3.5204	3.3125	3.1286	3.8530	0.651534	3.5062	0.008123	5.3580	3.1089
	60	1.9161	1.3255	1.2427	1.4485	0.25553	1.9106	0.002058	2.9082	1.8018

pling estimators are roughly equally effective. The IS/Rand estimator appears to outperform the IS/Reg estimator as ρ increases. The impact of the control variate becomes most dramatic in heavy traffic (the case studied by Asmussen 1990), precisely when the effectiveness of importance sampling seems to diminish.

5. Analysis of the Estimators

In this section, we prove Theorems 1–5. We begin with two lemmas. The following lemma affirms that the stopping times appearing in various algorithms are, in fact, almost surely finite.

LEMMA 2. *For all* $x > 0$,

(i) $\tilde{T}(x) = \inf\{n \geq 1: \tilde{S}_n^1 > x\}$ *is a.s. finite,*

(ii) $\tilde{T}'(x) = \inf\{n \geq 1: \tilde{S}_{n-1}^1 + \tilde{D}_n > x\}$ *is a.s. finite,*

(iii) $\tilde{\tau}(x) = \inf\{n \geq 1: \tilde{Y}_n^1 > x\}$ *is a.s. finite,*

(iv) $\tilde{\tau}'(x) = \inf\{n \geq 1: \tilde{Y}_{n-1}^1 + \tilde{D}_n > x\}$ *is a.s. finite, and*

(v) $\tilde{\tau}_0^d = \inf\{n \geq 1: \tilde{Y}_n^d = 0\}$ *is a.s. finite.*

Each of the assertions in Lemma 2 (i)–(iv) follows from the fact that a positive-drift random walk reaches any finite positive level in finite time, with probability 1 (Theorem III.1.1 of Gut 1988). The details are straightforward and therefore omitted. To argue the finiteness of $\tilde{\tau}_0^d$ appearing in (24), we consider these two cases. If the stopping time $\tilde{\tau}_0^d$ occurs before $\tilde{\tau}(s^1)$, the finiteness of the latter (as claimed in Lemma 2(iii)) certainly implies the finiteness of the former. If $\tilde{\tau}(s^1) < \tilde{\tau}_0^d$, the process $\{\tilde{Y}_n^d, \tilde{\tau}(s^1) < n < \tilde{\tau}_0^d\}$ is a negative-drift random walk, ensuring that the origin is reached in finite time for any s^1. The same argument applies for $\tilde{\tau}_0^d$ in (29).

Our next lemma, giving bounds on various quantities, is central to the upper bounds on the RE of our estimators. Define

$$C_- = \inf_{r \geq \epsilon}\{\mathsf{E}[e^{\gamma(D_1 - r)} \mid D_1 > r]\}^{-1}, \tag{30}$$

$$C_+ = \sup_{r \geq \epsilon}\{\mathsf{E}[e^{\gamma(D_1 - r)} \mid D_1 > r]\}^{-1}, \tag{31}$$

$$\zeta_- = \min_{n \geq 1}[r_n - nc^*], \quad \text{and} \tag{32}$$

$$\zeta_+ = \max_{n \geq 1}[r_n - nc^*], \tag{33}$$

in which $\epsilon \equiv \min[c^*, s^2 - s^1, \ldots, s^d - s^{d-1}]$.

LEMMA 3. *With the constants* C_-, C_+, ζ_-, *and* ζ_+ *above and stopping times* $\tilde{T}(x)$ *and* $\tilde{T}'(x)$ *defined in Lemma 2, we have for all* $x > 0$

(i) $C_- e^{-\gamma(x+\zeta_+)} \leq \mathsf{E}[e^{-\gamma \tilde{S}_{\tilde{T}(x)}}] \leq C_+ e^{-\gamma(x+\zeta_-)}$,

(ii) $\mathsf{E}[e^{-\gamma \tilde{S}_{\tilde{T}'(x)}} - e^{-\gamma \tilde{S}_{\tilde{T}(x)}}] \geq (1 - C_+)e^{-\gamma(x+\zeta_+)}$ *under condition* (20),

(iii) $\mathsf{E}[e^{-2\gamma \tilde{S}_{\tilde{T}(x)}}] \leq C_+ e^{-2\gamma(x+\zeta_-)}$, *and*

(iv) $\mathsf{E}[e^{-2\gamma \tilde{S}_{\tilde{T}'(x)}}] \leq C_+ e^{-2\gamma(x-c^*+\zeta_-)}$.

The proof of each case follows from an argument in Ross (1974). For instance, to derive (i) we condition on $\tilde{T}(x)$ and $\tilde{S}_{\tilde{T}(x)-1}$, use the renewal property of the ascending ladder heights of $\tilde{S}_n$, and apply Wald's likelihood ratio identity. A little algebraic simplification easily brings out the result. The remaining parts work similarly.

Now we are ready to turn to

PROOF OF THEOREM 1. Based on recursions (10)–(11) it is not hard to see $S_n^1 = \Sigma_1^n D_i - r_n$. Expanding (1)–(2) and using the i.i.d. property of demands gives rise to

$$Y_n^1 =_d \max_{0 \leq j \leq n}\left[\sum_1^j D_i - r_j\right] \quad \text{for } n \geq 0, \tag{34}$$

as in Lemma 2 of Glasserman (1993). Passing to the limit yields $Y^1 =_d \max_{n \geq 0}[\Sigma_1^n D_i - r_n] = \max_{n \geq 0} S_n^1$; that is, the steady-state shortfall Y^1 is equal in distribution to the maximum over all time of the sequence $\{S_n^1, n \geq 0\}$. If we define $T(s^1) = \inf\{n \geq 0: S_n^1 > s^1\}$, then $P\{Y^1 > s^1\} = P\{\max_{n \geq 0} S_n^1 > s^1\} = P\{T(s^1) < \infty\}$. An application of Wald's likelihood ratio identity gives $P\{T(s^1) < \infty\} = \mathsf{E}[e^{-\gamma \tilde{S}_{\tilde{T}(s^1)}}]$, establishing unbiasedness.

Because the RE of an estimator is the ratio of its standard error to its mean, it is bounded by the square root of the second moment divided by the mean. Thus, putting Lemma 3(i) and (iii) together, we have for all $s^1 > 0$

$$\mathrm{RE} \leq \frac{\sqrt{\mathsf{E}[e^{-2\gamma \tilde{S}_{\tilde{T}(s^1)}}]}}{\mathsf{E}[e^{-\gamma \tilde{S}_{\tilde{T}(s^1)}}]} \leq \frac{\sqrt{C_+ e^{-2\gamma(s^1+\zeta_-)}}}{C_- e^{-\gamma(s^1+\zeta_+)}} = \frac{\sqrt{C_+}}{C_-} e^{\gamma(\zeta_+ - \zeta_-)},$$

provided that $C_- > 0$. □

REMARK 4. (i) It has been pointed out that C_- could be zero in some exceptional cases. To circumvent the problem of a zero denominator, we make use of an asymptotic result developed in Glasserman (1993):

$$\alpha(s^1) = P\{Y^1 > s^1\} \sim Ce^{-\gamma(s^1+\eta)},$$

where η is defined in (18) and

$$C = \lim_{s^1 \to \infty} \mathrm{E}[e^{-\gamma(\hat{S}_{\hat{T}(s^1)} - s^1)}]. \tag{35}$$

We claim that

$$P\{Y^1 > s^1\} \geq (C/2)\exp[-\gamma(\max[s^1, s_0^1] + \eta)]$$

for some constant s_0^1. There always exists an s_0^1 such that for all $s^1 \geq s_0^1$,

$$P\{Y^1 > s^1\} \geq (C/2)e^{-\gamma(s^1+\eta)}.$$

On the other hand, for all $0 < s^1 < s_0^1$,

$$P\{Y^1 > s^1\} \geq P\{Y^1 > s_0^1\} \geq (C/2)e^{-\gamma(s_0^1+\eta)}.$$

Therefore, by properly incorporating this alternative lower bound, we find that the RE in Theorem 1 is bounded above by $2\sqrt{C_+}e^{\gamma(\eta-\zeta_-)}/C$ for $s^1 \geq s_0^1$ and by $2\sqrt{C_+}e^{\gamma(s_0^1+\eta-\zeta_-)}/C$ for $s^1 < s_0^1$. Similar modifications apply to our other bounds.

(ii) The constants C_-, C_+ are evaluated for some commonly used distributions (including Erlang and hyperexponential) in Glasserman (1993).

(iii) Combining the unbiasedness result $\mathrm{E}[\exp\{-\gamma\hat{S}_{\hat{T}(s^1)}\}] = \alpha(s^1)$ with part (i) of Lemma 3 results in bounds on the stockout probability $\alpha(s^1)$ in multistage systems. For a single-stage system (for which $\zeta_- = \zeta_+ = 0$) these bounds are precisely the ones given in Glasserman (1993). Corresponding bounds hold for $b(s^1)$ and $\beta(s^1)$.

Next we give

PROOF OF THEOREM 2. Because

$$\mathrm{E}\left[e^{-\gamma s^1}\int_{s^1}^{s^1+L} e^{-\gamma(\hat{S}_{\hat{T}(x)}-x)}dx\right]$$

$$= \mathrm{E}\left[\int_{s^1}^{\infty} e^{-\gamma\hat{S}_{\hat{T}(x)}}e^{\gamma(x-s^1)}\mathbf{1}_{\{L>x-s^1\}}dx\right]$$

$$= \int_{s^1}^{\infty} \mathrm{E}[e^{-\gamma\hat{S}_{\hat{T}(x)}}]e^{\gamma(x-s^1)}P\{L > x - s^1\}dx$$

$$= \int_{s^1}^{\infty} P\{Y^1 > x\}dx = b(s^1),$$

the estimator is unbiased. For the second assertion, we need to find an upper bound on the second moment of estimator (17). Since

$$\mathrm{E}\left[\left\{e^{-\gamma s^1}\int_{s^1}^{s^1+L} e^{-\gamma(\hat{S}_{\hat{T}(x)}-x)}dx\right\}^2\right]$$

$$= \mathrm{E}\left[\left\{\int_{s^1}^{\infty} e^{-\gamma\hat{S}_{\hat{T}(x)}}e^{\gamma(x-s^1)}\mathbf{1}_{\{L>x-s^1\}}dx\right\}^2\right]$$

$$\leq \mathrm{E}\left[\int_{s^1}^{\infty} e^{-2\gamma\hat{S}_{\hat{T}(x)}}e^{2\gamma(x-s^1)}\mathbf{1}_{\{L>x-s^1\}}dx\right]$$

$$= \int_{s^1}^{\infty} \{\mathrm{E}[e^{-2\gamma\hat{S}_{\hat{T}(x)}}]e^{2\gamma(x-s^1)}P\{L > x - s^1\}\}dx$$

$$= \int_{s^1}^{\infty} \{\mathrm{E}[e^{-2\gamma\hat{S}_{\hat{T}(x)}}]e^{\gamma(x-s^1)}\}dx,$$

it is bounded above by $\int_{s^1}^{\infty} C_+e^{-2\gamma(x+\zeta_-)}e^{\gamma(x-s^1)}dx$ according to Lemma 3(iii). On the other hand, (15) implies that the mean has a lower bound $\int_{s^1}^{\infty} C_-e^{-\gamma(x+\zeta_+)}dx$. It then follows that for all $s^1 > 0$,

$$\mathrm{RE} \leq \frac{\sqrt{\frac{C_+}{\gamma}}e^{-2\gamma(s^1+\zeta_-)}}{\frac{C_-}{\gamma}e^{-\gamma(s^1+\zeta_+)}} = \frac{\sqrt{\gamma C_+}}{C_-}e^{\gamma(\zeta_+-\zeta_-)}. \quad \square$$

We postpone our analysis of the regenerative case and next give

PROOF OF THEOREM 4. The unbiasedness is clear. To bound the RE of the estimator is essentially equivalent to finding a lower bound on its mean and an upper bound on its second moment. Lemma 3(ii) implies that the mean of estimator (27) is no less than $\int_{s^1}^{\infty}(1 - C_+)e^{-\gamma(x+\zeta_+)}dx$ if condition (20) is imposed. On the other hand, the second moment of (27) is bounded above by

$$e^{-2\gamma s^1}\mathrm{E}\left[\int_{s^1}^{s^1+L}\{e^{-\gamma(\hat{S}_{\hat{T}(x)}-x)} - e^{-\gamma(\hat{S}_{\hat{T}(x)}-x)}\}^2dx\right]$$

$$\leq e^{-2\gamma s^1}\mathrm{E}\left[\int_{s^1}^{s^1+L}\{e^{-2\gamma(\hat{S}_{\hat{T}(x)}-x)} + e^{-2\gamma(\hat{S}_{\hat{T}(x)}-x)}\}dx\right].$$

Again Lemma 3 is invoked to yield the upper bound

$$\frac{C_+}{\gamma}(e^{2\gamma c^*}+1)e^{-2\gamma(s^1+\zeta_-)}.$$

As a result, we have

$$\text{RE} \le \frac{\sqrt{\frac{C_+}{\gamma}(e^{2\gamma c^*}+1)e^{-2\gamma(s^1+\zeta_-)}}}{\frac{1-C_+}{\gamma}e^{-\gamma(s^1+\zeta_+)}}$$

$$= \frac{\sqrt{\gamma C_+(e^{2\gamma c^*}+1)}}{1-C_+}e^{\gamma(\zeta_+-\zeta_-)},$$

under condition (20). □

REMARK 5. Paralleling the asymptotic expression for $P\{Y^1 > s^1\}$ in Remark 4, we can also show that

$$P\{Y^1 + D > x\} \sim Ce^{-\gamma(x+\eta-c^*)},$$

with η and C given in (18) and (35), respectively. By (25)–(26) and (15) it follows that

$$(\mathsf{E}[D])(1-\beta(s^1)) \sim \int_{s^1}^{\infty} C(e^{\gamma c^*}-1)e^{-\gamma(x+\eta)}dx$$

$$= \frac{C}{\gamma}(e^{\gamma c^*}-1)e^{-\gamma(s^1+\gamma)}.$$

Again, we assert that

$$(\mathsf{E}[D])(1-\beta(s^1)) \ge \frac{C}{2\gamma}(e^{\gamma c^*}-1)e^{-\gamma(\max[s^1,s^1_*]+\eta)},$$

for some constant s^1_*. Therefore, without (20) we still have a lower bound on $(\mathsf{E}[D])(1-\beta(s^1))$, and therefore an upper bound on the RE of the estimator (27).

We next consider the regenerative estimators (24) and (29), beginning with

PROOF OF LEMMA 1. Let empty sums be zero and let $a \vee b$ and $a \wedge b$ denote $\max\{a, b\}$ and $\min\{a, b\}$, respectively. Expanding the shortfalls in recursions (1)–(2) gives

$$Y_n^d = \max_{0\le j\le n}\left\{\sum_{i=n+1-j}^{n} D_i - jc^d\right\} \equiv \sum_{i=n+1-l}^{n} D_i - lc^d;$$

$$Y_n^k = \max_{0\le j\le n}\left\{\sum_{i=n+1-j}^{n} D_i - r_j^k\right\} \equiv \sum_{i=n+1-h}^{n} D_i - r_h^k.$$

The random integers $0 \le l, h \le n$ are indices at which each maximum is achieved. There are three possible cases.

If $l, h > 0$ (where l and h may or may not be the same), then

$$Y_n^d - Y_n^k \le \left(\sum_{i=n+1-l}^{n} D_i - lc^d\right) - \left(\sum_{i=n+1-l}^{n} D_i - r_l^k\right)$$

$$\le \max_{1\le j\le n}[r_j^k - jc^d].$$

If $l = 0, h > 0$, then

$$Y_n^d - Y_n^k \le 0 - \left(\sum_{i=n+1-h}^{n} D_i - r_h^k\right) \le 0.$$

If $l > 0, h = 0$, then

$$Y_n^d - Y_n^k \le \left(\sum_{i=n+1-l}^{n} D_i - lc^d\right) - \left(\sum_{i=n+1-l}^{n} D_i - r_l^k\right)$$

$$\le \max_{1\le j\le n}[r_j^k - jc^d].$$

We then conclude that

$$Y_n^d - Y_n^k \le \max_{1\le j\le n}[r_j^k - jc^d] \vee 0. \tag{36}$$

On the other hand, if the roles of Y_n^d and Y_n^k are reversed, we get

$$Y_n^k - Y_n^d \le \max_{1\le j\le n}[jc^d - r_j^k] \vee 0. \tag{37}$$

Multiplying each side of (37) by -1 and combining it with (36) yield

$$\min_{1\le j\le n}[r_j^k - jc^d] \wedge 0 \le Y_n^d - Y_n^k \le \max_{1\le j\le n}[r_j^k - jc^d] \vee 0,$$

which is the claim in part (i). Part (ii) follows from part (i). □

Our next lemma is useful in bounding the excess of shortfalls over some level during a regenerative cycle.

LEMMA 4. *Let the random variable Z satisfy $Z > u$ a.s., for some constant $u > 0$, and let the i.i.d. random variables $\{X_n, n \ge 1\}$ be independent of Z and have mean $\mathsf{E}[X_1] < 0$. Suppose there exists a $\gamma > 0$ for which $\mathsf{E}[e^{\gamma X_1}] = 1$. Define $\mathbf{t} = \inf\{n \ge 1: Z + X_1 + \cdots + X_n \le 0\}$. Then*

$$\mathsf{E}\left[\left\{\sum_{n=0}^{t-1}(Z+X_1+\cdots+X_n-u)^+\right\}^2\right]$$
$$\le \frac{32}{(\mathsf{E}[X_1])^2}\left\{\mathsf{E}\left[\left(Z-u-\frac{\mathsf{E}[X_1]}{2}\right)^4\right]+\frac{24}{\hat{\gamma}^4}\right\},$$

where $\hat{\gamma}$ *solves* $\mathsf{E}[\exp\{\hat{\gamma}(X_1-\mathsf{E}[X_1]/2)\}]=1$.

PROOF. Define $N_y=\max\{n\ge 1: X_1+\cdots+X_n\ge -y\}$ for $y>0$. That is, for a negative-drift random walk starting from the origin, N_y is the last-exit time over level $-y$. Then, by applying an argument in Janson (1986), we have

$$\frac{\mathsf{E}[X_1]}{2}N_y \ge \frac{\mathsf{E}[X_1]}{2}N_y-(X_1+\cdots+X_{N_y}+y)$$
$$=-y-\sum_{i=1}^{N_y}\left(X_i-\frac{\mathsf{E}[X_1]}{2}\right)$$
$$\ge -y-\max_{n\ge 0}\left[\sum_{i=1}^{n}\left(X_i-\frac{\mathsf{E}[X_1]}{2}\right)\right]. \quad (38)$$

The random variable $M'\equiv\max_{n\ge0}[\Sigma_{i=1}^{n}(X_i-(\mathsf{E}[X_1]/2))]$ is a.s. finite since the increment $X_i-(\mathsf{E}[X_1]/2)$ has negative mean. Note that $X_1-(\mathsf{E}[X_1]/2)\ge X_1$ a.s., so

$$0\le M\le M'<\infty \quad \text{a.s.}, \quad (39)$$

with $M\equiv\max_{n\ge0}[\Sigma_{i=1}^{n}X_i]$. It then follows from (38) that

$$N_y\le\frac{2}{-\mathsf{E}[X_1]}(y+M') \quad \text{a.s. for all } y>0. \quad (40)$$

Viewing $\Sigma_{n=0}^{t-1}(Z+X_1+\cdots+X_n-u)^+$ as the area above level u under a piecewise constant interpolation of the path of a random walk, it becomes evident that this quantity can be no greater than the product of $N_{Z-u}+1$ and the maximum height $Z-u+M$, i.e.,

$$\mathsf{E}\left[\left\{\sum_{n=0}^{t-1}(Z+X_1+\cdots+X_n-u)^+\right\}^2\right]$$
$$\le \mathsf{E}[(N_{Z-u}+1)^2(Z-u+M)^2]. \quad (41)$$

By conditioning on $Z-u$, we have

$$\mathsf{E}[(N_y+1)^2(y+M)^2\,|\,Z-u=y]$$
$$\le\sqrt{\mathsf{E}[(N_y+1)^4\,|\,Z-u=y]}\sqrt{\mathsf{E}[(y+M)^4]}$$
(by Cauchy-Schwarz Inequality and independence of M and Z)
$$\le\sqrt{\frac{16}{(\mathsf{E}[X_1])^4}\mathsf{E}\left[\left(y+M'-\frac{\mathsf{E}[X_1]}{2}\right)^4\,|\,Z-u=y\right]}$$
$$\times\sqrt{\mathsf{E}[(y+M)^4]} \quad \text{(by (40))}$$
$$\le\sqrt{\frac{16}{(\mathsf{E}[X_1])^4}\mathsf{E}\left[\left(y+M'-\frac{\mathsf{E}[X_1]}{2}\right)^4\right]}\sqrt{\mathsf{E}[(y+M)^4]}$$
(by independence of M' and Z)
$$\le\frac{4}{(\mathsf{E}[X_1])^2}\mathsf{E}\left[\left(y+M'-\frac{\mathsf{E}[X_1]}{2}\right)^4\right) \quad \text{(by (39))}$$
$$\le\frac{32}{(\mathsf{E}[X_1])^2}\left\{\mathsf{E}\left[\left(y-\frac{\mathsf{E}[X_1]}{2}\right)^4\right]+\mathsf{E}[(M')^4]\right\}.$$

The last expression is an immediate consequence of the inequality

$$(a+b)^p\le\max(2^{p-1},1)(a^p+b^p) \quad \text{for } a,b,p>0, \quad (42)$$

which is taken from Exercise 4.2.1, Chow and Teicher (1988). Hence (41) is bounded above by

$$\frac{32}{(\mathsf{E}[X_1])^2}\left\{\mathsf{E}\left[\left(Z-u-\frac{\mathsf{E}[X_1]}{2}\right)^4\right]+\mathsf{E}[(M')^4]\right\}. \quad (43)$$

On the other hand, by (5) we know that

$$\mathsf{E}\left[\exp\left\{\gamma\left(X_1-\frac{\mathsf{E}[X_1]}{2}\right)\right\}\right]=\exp\left\{\gamma\left(-\frac{\mathsf{E}[X_1]}{2}\right)\right\}>1,$$

which ensures the existence of a positive $\hat{\gamma}$ at which

$$\mathsf{E}\left[\exp\left\{\hat{\gamma}\left(X_1-\frac{\mathsf{E}[X_1]}{2}\right)\right\}\right]=1.$$

By quoting a well-known result that

$$P\{M'>x\}<e^{-\hat{\gamma}x},$$

(see, e.g., p. 269 of Asmussen 1987) we get

$$\mathsf{E}[(M')^4] = 4\int_0^\infty x^3 P\{M' > x\}dx$$

$$\le 4\int_0^\infty x^3 e^{-\hat{\gamma}x}dx = \frac{24}{\hat{\gamma}^4}.$$

This, together with (43), proves the claim.

PROOF OF THEOREM 3. According to the algorithm, we use new demands $\tilde{D}_n$ up until $\tilde{\tau}(s^1)$, so the likelihood ratio is simply $e^{-\gamma \tilde{S}_{\tilde{\tau}(s^1)}}$. Also, by definition of $\tilde{\tau}(s^1)$, $\tilde{Y}_n^1 \le s^1$ for all $n < \tilde{\tau}(s^1)$. The estimator is thus unbiased, by Wald's likelihood ratio identity. For the second assertion, we observe that Assumption (19) makes the conjugate random walk $\{\tilde{S}_n,\ n \ge 0\}$ coincide with the process $\{\tilde{Y}_n^d,\ n \ge 0\}$ prior to T_1. Also, we observe that Lemma 1(i)–(ii) implies $\tilde{Y}_n^1 \le \tilde{Y}_n^d \le \tilde{Y}_n^1 + [\zeta_+]^+$ for all $n \ge 0$. Thus, $\tilde{Y}^d$ cannot cross level $s^1 + [\zeta_+]^+$ before $\tilde{Y}^1$ crosses level s^1. Recall that $\tilde{\tau}(s^1)$ given in (23) denotes the first time $\tilde{Y}^1$ exceeds level s^1. If $\tilde{Y}^d$ first crosses level $s^1 + [\zeta_+]^+$ at $\tilde{\tau}(s^1)$, we know that $\tilde{Y}^d_{\tilde{\tau}(s^1)} = s^1 + [\zeta_+]^+ + R_{s^1+[\zeta_+]^+}$ where $R_{s^1+[\zeta_+]^+}$ is the excess of $\tilde{S}_{\tilde{\tau}(s^1)}$ over $s^1 + [\zeta_+]^+$. Otherwise, $\tilde{Y}^d_{\tilde{\tau}(s^1)} \le s^1 + [\zeta_+]^+$ if $\tilde{Y}^d$ has not yet reached level $s^1 + [\zeta_+]^+$ at $\tilde{\tau}(s^1)$. In any case, the relation

$$\tilde{S}_{\tilde{\tau}(s^1)} = \tilde{Y}^d_{\tilde{\tau}(s^1)} \le s^1 + [\zeta_+]^+ + R_{s^1+[\zeta_+]^+}, \qquad (44)$$

holds on the set $\{\tilde{\tau}(s^1) < \tilde{\tau}_0^d\}$. Therefore, the second moment of (24) is bounded above by

$$e^{-2\gamma s^1}\mathsf{E}\left[\left\{\sum_{n=\tilde{\tau}(s^1)}^{\tilde{\tau}_0^d-1} (\tilde{Y}_n^1 - s^1)^+ \mathbf{1}_{\{\tilde{\tau}(s^1)<\tilde{\tau}_0^d\}}\right\}^2\right] \quad (\text{since } \tilde{S}_{\tilde{\tau}(s^1)} = \tilde{Y}^d_{\tilde{\tau}(s^1)} \ge \tilde{Y}^1_{\tilde{\tau}(s^1)} > s^1)$$

$$\le e^{-2\gamma s^1}\mathsf{E}\left[\left\{\sum_{n=\tilde{\tau}(s^1)}^{\tilde{\tau}_0^d-1} (\tilde{Y}_n^d - s^1)^+ \mathbf{1}_{\{\tilde{\tau}(s^1)<\tilde{\tau}_0^d\}}\right\}^2\right] \quad (\text{by Lemma 1(ii)})$$

$$= e^{-2\gamma s^1}\mathsf{E}\left[\left\{\sum_{n=\tilde{\tau}(s^1)}^{\tilde{\tau}_0^d-1} (\tilde{S}_{\tilde{\tau}(s^1)} + X_{\tilde{\tau}(s^1)+1} + \cdots + X_n - s^1)^+ \mathbf{1}_{\{\tilde{\tau}(s^1)<\tilde{\tau}_0^d\}}\right\}^2\right]$$

(since $\tilde{Y}_n^d = \tilde{S}_{\tilde{\tau}(s^1)} + X_{\tilde{\tau}(s^1)+1} + \cdots + X_n$ for $\tilde{\tau}(s^1) \le n < \tilde{\tau}_0^d$)

$$= e^{-2\gamma s^1}\mathsf{E}\left[\left\{\sum_{n=0}^{\tilde{\tau}_0^d-\tilde{\tau}(s^1)-1} (\tilde{S}_{\tilde{\tau}(s^1)} + X_1 + \cdots + X_n - s^1)^+ \mathbf{1}_{\{\tilde{\tau}(s^1)<\tilde{\tau}_0^d\}}\right\}^2\right]$$

$$\le e^{-2\gamma s^1}\frac{32}{(\mathsf{E}[X_1])^2}\left\{\mathsf{E}\left[\left(\tilde{S}_{\tilde{\tau}(s^1)} - s^1 - \frac{\mathsf{E}[X_1]}{2}\right)^4\right] + \frac{24}{\hat{\gamma}^4}\right\} \quad (\text{see explanation below})$$

$$\le e^{-2\gamma s^1}\frac{32}{(\mathsf{E}[X_1])^2}\left\{\mathsf{E}\left[\left([\zeta_+]^+ + R_{s^1+[\zeta_+]^+} - \frac{\mathsf{E}[X_1]}{2}\right)^4\right] + \frac{24}{\hat{\gamma}^4}\right\} \quad (\text{by (44)})$$

$$\le e^{-2\gamma s^1}\frac{32}{(\mathsf{E}[X_1])^2}\left\{8\left([\zeta_+]^+ - \frac{\mathsf{E}[X_1]}{2}\right)^4 + \sup_{s^1\ge 0}\mathsf{E}[R^4_{s^1+[\zeta_+]^+}]\right) + \frac{24}{\hat{\gamma}^4}\right\} \quad (\text{by (42)})$$

$$\le e^{-2\gamma s^1}\frac{32}{(\mathsf{E}[X_1])^2}\left\{8\left([\zeta_+]^+ - \frac{\mathsf{E}[X_1]}{2}\right)^4 + \frac{48}{5}\frac{\mathsf{E}[(\tilde{X}_1^-)^5]}{\mathsf{E}[\tilde{X}_1]} + \frac{24}{\hat{\gamma}^4}\right\} \quad (\text{by Theorem 3, Lorden (1970)}). \qquad (45)$$

The notation $\tilde{X}_1^-$ is short for $-\min\{\tilde{X}_1, 0\}$, the negative part of $\tilde{X}_1$. On the event $\{\tilde{\tau}(s^1) < \tilde{\tau}_0^d\}$, the term $(\tilde{\tau}_0^d - \tilde{\tau}(s^1))$ in (45) coincides with the first time the process $\{\tilde{S}_{\tilde{\tau}(s^1)} + X_1 + \cdots + X_n\}$ takes a value less

than or equal to zero. This, combined with the relation $\tilde{S}_{\tilde{\tau}'(s^1)} > s^1$, makes Lemma 4 applicable. By setting

$$B = \frac{32}{(\mathrm{E}[X_1])^2} \times \left\{ 8\left([\zeta_+]^+ - \frac{\mathrm{E}[X_1]}{2} \right)^4 + \frac{48}{5} \frac{\mathrm{E}[(\hat{X}_1^-)^5]}{\mathrm{E}[\hat{X}_1]} + \frac{24}{\hat{\gamma}^4} \right\}, \quad (46)$$

and using a lower bound on the first moment of (24), we thus find that

$$\mathrm{RE} \le \frac{\sqrt{e^{-2\gamma s^1} B}}{\frac{C_-}{\gamma} e^{-\gamma(s^1+\zeta_+)}} = \frac{\gamma e^{\gamma\zeta_+}\sqrt{B}}{C_-}.$$

The fact that $\mathrm{E}[\hat{X}_1^5]$ is finite, noted in the last paragraph in §3, removes the possibility that this upper bound is infinite. □

PROOF OF THEOREM 5. The unbiasedness is obvious. Next, by observing $\tilde{Y}_n^d \le \tilde{Y}_{n-1}^1 + \tilde{D}_n - c^d + [\zeta_+]^+$ for $1 \le n \le T_2$ and paralleling the argument leading to (44), we conclude that

$$\tilde{S}_{\tilde{\tau}'(s^1)} - (s^1 - c^d) \le [\zeta_+]^+ + R_{s^1 - c^d + [\zeta_+]^+}, \quad (47)$$

on the set $\{\tilde{\tau}'(s^1) < \tilde{\tau}_0^d\}$, where $R_{s^1-c^d+[\zeta_+]^+}$ is the overshoot of $\tilde{S}_{\tilde{\tau}'(s^1)}$ over level $s^1 - c^d + [\zeta_+]^+$. Therefore, the upper bound on the second moment of estimator (29) can be brought out the same way. The second moment of estimator (29) is bounded by

$$\begin{aligned}
& e^{-2\gamma(s^1-c^d)}\mathrm{E}\left[\left\{\sum_{n=\tilde{\tau}'(s^1)}^{\tilde{\tau}_0^d-1} [(\tilde{Y}_{n-1}^1 + D_n - s^1)^+ \wedge D_n]\mathbf{1}_{\{\tilde{\tau}'(s^1)<\tilde{\tau}_0^d\}}\right\}^2\right] \\
& \quad \text{(since } \tilde{S}_{\tilde{\tau}'(s^1)} = \tilde{Y}_{\tilde{\tau}'(s^1)}^d \ge \tilde{Y}_{\tilde{\tau}'(s^1)-1}^1 + \tilde{D}_{\tilde{\tau}'(s^1)} - c^d > s^1 - c^d) \\
& \le e^{-2\gamma(s^1-c^d)}\mathrm{E}\left[\left\{\sum_{n=\tilde{\tau}'(s^1)}^{\tilde{\tau}_0^d-1} (\tilde{Y}_n^d + c^d - s^1)^+\mathbf{1}_{\{\tilde{\tau}'(s^1)<\tilde{\tau}_0^d\}}\right\}^2\right] \\
& \quad \text{(since } \tilde{Y}_{n-1}^1 + D_n - c^d \le \tilde{Y}_n^d \text{ for } n \ge \tilde{\tau}'(s^1)) \\
& = e^{-2\gamma(s^1-c^d)}\mathrm{E}\left[\left\{\sum_{n=\tilde{\tau}'(s^1)}^{\tilde{\tau}_0^d-1} [\tilde{S}_{\tilde{\tau}'(s^1)} + X_{\tilde{\tau}'(s^1)+1} + \cdots + X_n - (s^1 - c^d)]^+\mathbf{1}_{\{\tilde{\tau}'(s^1)<\tilde{\tau}_0^d\}}\right\}^2\right] \\
& \quad \text{(since } \tilde{Y}_n^d = \tilde{S}_{\tilde{\tau}'(s^1)} + X_{\tilde{\tau}'(s^1)+1} + \cdots + X_n \text{ for } \tilde{\tau}'(s^1) \le n < \tilde{\tau}_0^d) \\
& = e^{-2\gamma(s^1-c^d)}\mathrm{E}\left[\left\{\sum_{n=0}^{\tilde{\tau}_0^d-\tilde{\tau}'(s^1)-1} [\tilde{S}_{\tilde{\tau}'(s^1)} + X_1 + \cdots + X_n - (s^1 - c^d)]^+\mathbf{1}_{\{\tilde{\tau}'(s^1)<\tilde{\tau}_0^d\}}\right\}^2\right] \\
& \le e^{-2\gamma(s^1-c^d)} \frac{32}{(\mathrm{E}[X_1])^2}\left\{\mathrm{E}\left[\left(\tilde{S}_{\tilde{\tau}'(s^1)} - (s^1 - c^d) - \frac{\mathrm{E}[X_1]}{2}\right)^4\right] + \frac{24}{\hat{\gamma}^4}\right\} \quad \text{(by Lemma 4)} \\
& \le e^{-2\gamma(s^1-c^d)} \frac{32}{(\mathrm{E}[X_1])^2}\left\{\mathrm{E}\left[\left([\zeta_+]^+ + R_{s^1-c^d+[\zeta_+]^+} - \frac{\mathrm{E}[X_1]}{2}\right)^4\right] + \frac{24}{\hat{\gamma}^4}\right\} \quad \text{(by (47))}.
\end{aligned}$$

With B as in (46) we have the upper bound $Be^{-2\gamma(s^1-c^d)}$. By virtue of unbiasedness we also know that the mean of the estimator (29) has a lower bound $(1 - C_+/\gamma)e^{-\gamma(s^1+\zeta_+)}$. Finally we obtain

$$\mathrm{RE} \le \frac{\gamma\sqrt{B}}{1 - C_+} e^{\gamma(\zeta_+ + c^d)}. \quad \square$$

[1] This work was supported, in part, by NSF grants MSS-9216490 and DMI-94-57189. The authors thank the referees for their detailed comments.

References

Asmussen, S., *Applied Probability and Queues*, Wiley, Chichester, England, 1987.

——, "Exponential Families and Regression in the Monte Carlo Study of Queues and Random Walks," *Ann. Statistics*, 18 (1990), 1851–1867.

Chang, C. S., P. Heidelberger, S. Juneja, and P. Shahabuddin, "Effective Bandwidth and Fast Simulation of ATM Intree Networks," *Perf. Eval.*, 20 (1994), 45–65.

Chow, Y. S. and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*, Springer-Verlag, New York, 1988.

Clark, A. J. and H. Scarf, "Optimal Policies for a Multi-Echelon Inventory Problem," *Management Sci.*, 6 (1960), 475–490.

Federgruen, A. and P. Zipkin, "An Inventory Model with Limited Production Capacity and Uncertain Demands, I: The Average Cost Criterion," *Math. Oper. Res.*, 11 (1986), 193–207, "II: The Discounted Cost Criterion," 208–215.

Glasserman, P. and S. Tayur, "The Stability of a Capacitated, Multi-Echelon Production Inventory System Under a Base-Stock Policy," *Oper. Res.*, 42 (1994), 913–925.

——, "Bounds and Asymptotics for Planning Critical Safety Stocks," Working Paper, Columbia University, New York, 1993. To appear in *Oper. Res.*

Goyal, A., P. Shahabuddin, P. Heidelberger, V. F. Nicola, and P. W. Glynn, "A Unified Framework for Simulating Markovian Models of Highly Reliable Systems," *IEEE Trans. Computers*, C-41 (1992), 36–51.

Gut, A., *Stopped Random Walks*, Springer, New York, 1988.

Heidelberger, P., "Fast Simulation of Rare Events in Queueing and Reliability Models," *ACM Trans. Modeling and Computer Simulation*, 5 (1995), 43–85.

Janson, S., "Moments for First-Passage and Last-Exit Times, the Minimum, and Related Quantities for Random Walks with Positive Drift," *Adv. Appl. Prob.*, 18 (1986), 865–879.

Langenhoff, L. J. G. and W. H. M. Zijm, "An Analytical Theory of Multi-Stage Production / Distribution Systems," *Statistica Neerlandica*, 44 (1992), 149–174.

Lehtonen, T. and H. Nyrhinen, "Simulating Level-Crossing Probabilities by Importance Sampling," *Adv. Appl. Prob.*, 24 (1992), 858–874.

Lorden, G., "On Excess over the Boundary," *Ann. Math. Statist.*, 41 (1970), 520–527.

Nakayama, M., "Efficient Methods for Generating Some Exponentially Tilted Random Variates," in J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson (Eds.), *Proc. 1992 Winter Simulation Conf.*, Society for Computer Simulation, San Diego, CA, 1992, 603–608.

Rosling, K., "Optimal Inventory Policies for Assembly Systems Under Random Demands," *Oper. Res.*, 37 (1989), 565–579.

Ross, S. M., "Bounds on the Delay Distribution in GI/G/1 Queues," *J. Appl. Prob.*, 11 (1974), 417–421.

Sadowsky, J. S., "Large Deviations and Efficient Simulation of Excessive Backlogs in a GI/G/m Queue," *IEEE Trans. Automatic Control*, 36 (1991), 1383–1394.

Shahabuddin, P., "Importance Sampling for the Simulation of Highly Reliable Markovian Systems," *Management Sci.*, 40 (1994), 333–352.

Siegmund, D., "Importance Sampling in the Monte Carlo Study of Sequential Tests," *Ann. Statistics*, 4 (1976), 673–684.

Tayur, S., "Computing the Optimal Policy for Capacitated Inventory Models," *Stochastic Models*, 9 (1993), 585–598.

Accepted by Pierre L'Ecuyer; received May 12, 1994. *This paper has been with the authors* 6 *months for* 2 *revisions.*