A Large Deviations Perspective on the Efficiency of Multilevel Splitting

Paul Glasserman, Philip Heidelberger, *Fellow, IEEE*, Perwez Shahabuddin, *Member, IEEE*, and Tim Zajic

***Abstract*—Stringent performance standards for computing and telecommunications systems have motivated the development of efficient techniques for estimating rare event probabilities. In this paper, we analyze the performance of a multilevel splitting method for rare event simulation related to one recently proposed in the telecommunications literature. This method splits promising paths into subpaths at intermediate levels to increase the number of observations of a rare event. In a previous paper we gave sufficient conditions, in specific classes of models, for this method to be *asymptotically optimal*; here we focus on necessary conditions in a general setting. We show, through a variety of results, the importance of choosing the intermediate thresholds in a way consistent with the most likely path to a rare set, both when the number of levels is fixed and when it increases with the rarity of the event. In the latter case, we give very general necessary conditions based on large deviations rate functions. These indicate that even when the intermediate levels are chosen appropriately, the method will frequently fail to be asymptotically optimal. We illustrate the conditions with examples.**

***Index Terms*—Large deviations, Monte Carlo, rare event, simulation, variance reduction.**

I. Introduction

A. Background and Summary

DEVELOPMENTS in computing and telecommunications technology over roughly the last decade have brought special significance to rare events, and this in turn has driven the development of new modeling and analysis tools. Performance standards for failure probabilities in fault-tolerant computing and buffer overflow probabilities in ATM networks, for example, are stringent enough to make rare event asymptotics relevant for performance analysis. This has led to a burgeoning literature on methods based on large deviations techniques in particular; see, e.g., [3], [4], [19], [30], [33], and the references there.

Approximations based on asymptotics must ordinarily be supplemented with simulation for a more precise analysis, but the estimation of rare event probabilities by simulation poses a serious computational challenge. Indeed, for sufficiently small probabilities, straightforward simulation is simply infeasible and the use of powerful variance reduction techniques becomes essential. Interestingly, based in part on early work of Cottrell *et al.* [8], Parekh and Walrand [24], Siegmund [31], and others, it is now well known that the same large deviations asymptotics that give a rough approximation to a rare event probability often suggest a highly effective *change of measure* for variance reduction via *importance sampling* (see, e.g., [3], [5], [6], [12], [21], [23], [25], [26], and [29]).

This observation has led to numerous successful implementations of importance sampling in computer and communications applications and, furthermore, to an interesting link between optimal simulation and effective bandwidths [5], [23], [33]. At the same time, the method faces two serious shortcomings. First, it requires that the model to be simulated be amenable to a large deviations analysis, and this currently excludes most networks. Second, and perhaps even more worrisome, results in [16], [17], and [27] show that importance sampling techniques suggested by large deviations are not automatically effective and may in fact lead to poor results. These references show how extremely unlikely sample paths can contribute significantly to the variance of the estimate. This can lead to a loss in efficiency and, in some cases, can even result in an infinite variance. Thus importance sampling should be applied with caution.

Against this backdrop, a much simpler approach to rare event simulation recently advanced in a series of papers focused on telecommunications looks attractive. The method, introduced in Villén-Altamirano and Villén-Altamirano [35] and called RESTART there and in [28], [34], and [36], is in fact a multilevel *splitting* technique of the type used in simulation at least since Kahn and Harris [22] and frequently used in physics applications in particular (e.g., [11] and [32]). The technique is illustrated in Fig. 1. Suppose we want to estimate the probability that a process reaches a rare set A before returning to the origin, starting from the origin. (Estimation of steady-state rare events like buffer overflow probabilities and failure probabilities can often be reduced to the estimation of this type of probability; see [19] and [24].) Since few paths make it to A, straightforward simulation may require a very large number of trials to produce an estimate of reasonable accuracy. Multilevel splitting introduces a series of intermediate thresholds between the starting state zero and the target set A; in the figure there are just two, but in practice there could be many. Each path that reaches a threshold is split

Manuscript received December 19, 1996; revised December 1, 1997. Recommended by Associate Editor, W.-B. Gong. This work was supported by the NSF National Young Investigator Award under Grant DMI-94-57189, the NSF University-Industry Cooperative Research Program under Grant DMS-95-08709, and the NSF Career Award under Grant DMI-96-25297.

P. Glasserman is with the Graduate School of Business, Columbia University, New York, NY 10027 USA.

P. Heidelberger is with IBM, T. J. Watson Research Center, Yorktown Heights, NY 10598 USA.

P. Shahabuddin is with the IEOR Department, Columbia University, New York, NY 10027 USA.

T. Zajic is with the School of Mathematics, University of Minnesota, Minneapolis, MN 55455 USA.

Publisher Item Identifier S 0018-9286(98)09437-9.

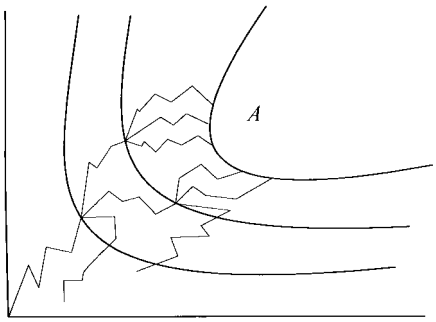


Fig. 1. Illustration of multilevel splitting with two intermediate levels and a splitting parameter of three. Of the three subpaths at the first split, two reach the next level, and of their subpaths, all but one reach A.

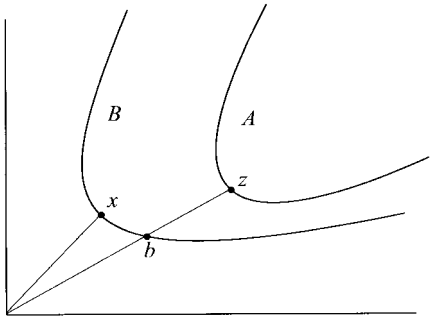


Fig. 2. The most likely path to B ends at x, but the most likely path to A hits B at b.

into a (possibly level-dependent) number of subpaths which subsequently evolve independently. Each path is terminated upon entry into A or zero. Dividing the number of paths that reach A by the product of the splitting parameters yields an unbiased estimate of the probability of reaching A before zero.

Much of the attractiveness of this method lies in its simplicity. It does not appear to require an extensive analysis of the underlying process for implementation; indeed it appears to be almost universally applicable. Appropriately implemented, it promises to use computation time effectively by reinforcing the informative paths that reach intermediate thresholds.

In [13] we identified a class of systems and implementation conditions for which multilevel splitting results in *asymptotically optimal* estimates of rare event probabilities. This notion is reviewed in Section IV; briefly, it ensures that the computational effort required to achieve a fixed precision does not grow too quickly with the rarity of the event. This is the standard criterion in theoretical analyses of importance sampling for large deviations rare events. Hence, the results in [13] show that—under the right conditions—multilevel splitting is as asymptotically effective as the best importance sampling estimators arrived at using large deviations techniques. A key observation in [13] is that the splitting parameters should be chosen to keep the expected number of surviving subpaths at each level roughly constant. This balances the loss of variance reduction from too little splitting and the exponential growth in computational effort from too much splitting.

The purpose of this paper is to explore in more detail the relation between effective splitting and the large deviations behavior of the underlying process. We show that, despite its apparent simplicity, splitting ultimately relies on a detailed understanding of a process's rare event asymptotics, much as importance sampling does. This leads to necessary conditions for asymptotic optimality that balance the sufficient conditions given in [13] and show that splitting is by no means a panacea for rare event simulation.

A central issue in implementing multilevel splitting is the choice of thresholds. A little thought suggests that they should be chosen consistent (in some sense) with the most likely path to the rare set—the path sought by a large deviations analysis. To see why, consider the setting illustrated in Fig. 2. Suppose that, conditional on reaching A before zero, the process tends to follow the path ending at z, and conditional on reaching B before zero it tends to follow the path ending at x. Since the paths do not coincide, we would say that the intermediate threshold has not been chosen consistent with the most likely path.

If we now apply splitting in this setting, we will end up reinforcing a lot of subpaths that start near x. To reach A it would be better to have many subpaths starting near b, but few of the paths that make it to B will hit B near b. This suggests that unless we know how to choose the intermediate thresholds to make the conditional entry points (nearly) coincide, splitting will not result in an effective allocation of computational effort. Put another way, the most likely path to an intermediate level must coincide with the most likely path to the final level.

This insight will be made precise in various ways throughout the rest of the paper. In Section II, we treat cases with a fixed number of levels and small probabilities of moving from one level to the next; this type of setting arises in models of highly reliable computing systems. We show that under conditions corresponding to Fig. 2, there is indeed a loss of efficiency. Furthermore, in Section III, we show that there is a type of instability resulting from the mismatch between the entrance distributions at B conditional on hitting A or B; with high probability, the splitting estimator will appear biased over even a large number of runs. A similar problem can occur in importance sampling if the simulation is "overbiased" [1], [10], [26]. In the context of importance sampling, this phenomenon was called "apparent bias" in [10]; we adopt this terminolgy when this phenomenon occurs in the context of splitting.

In Section IV we develop necessary conditions for asymptotic efficiency in cases where events become rare because the number of levels increases while the dynamics of the underlying process remain fixed; this setting is relevant to buffer overflow models, for example. We first give a necessary condition on the splitting parameter: the log of this parameter should equal the exponential rate of decay of the probability to be estimated. This is consistent with *sufficient* conditions given in [13] under much more specific assumptions *and implies that the expected number of subpaths entering each level neither grows nor shrinks too quickly*.

Next we show that even choosing the right splitting parameter does not in general guarantee asymptotic optimality (though it did in the special cases analyzed in [13]). The problem in the general case again arises from the possibility illustrated in

Fig. 2. To preclude this possibility in a very strong sense, we give a further necessary condition based on large deviations techniques. This condition may be interpreted as requiring a strong tendency to follow the most likely path, conditional on reaching a rare set. We give a class of processes for which this necessary condition is satisfied.

Section V shows that our necessary conditions are violated by a simple Jackson network under what appears to be the optimal choice of thresholds, suggesting that even if splitting is implemented in the best possible way it may fail to be asymptotically efficient. Numerical results confirm that splitting fails to be completely effective in this Jackson network example. A brief discussion summarizing the implications of the results in [13] and this paper may be found in [15].

B. Framework and Notation

To obtain reasonable generality without excessive complexity, we assume throughout that the underlying process to be simulated is a countable-state Markov chain in either discrete or continuous time. The initial distribution of the process is concentrated on a set of states A_0 to which the process returns with probability one after exiting. We are interested in estimating γ, the probability of hitting a target set A before returning to A_0. To apply multilevel splitting, we define a nested sequence of sets (also called levels), $A_1, A_2, \cdots, A_{k-1}$, with $A_1 \supset A_2 \supset \cdots \supset A_{k-1} \supset A_k \equiv A$ and $A_1 \cap A_0 = \emptyset$. We assume that the process cannot enter A_{i+1}, $i \geq 1$, without first entering $A_{i+1}^c \cap A_i$, where A_{i+1}^c is the complement of A_{i+1}. Paths are split upon *first* entrances to the sets A_i. We can think of the process being absorbed in A_0, when it reenters A_0. Hence, there may be a positive probability that a set A_i, $i \geq 1$ is never entered. Let p_i denote the probability of entering the set A_i, given that the set A_{i-1} has been entered. We set $\gamma_i = \prod_{j=1}^{i} p_j$ and then $\gamma \equiv \gamma_k$. We assume that each p_i is positive, as otherwise the problem is trivial.

Let R_i denote the splitting parameter for level $(i-1)$—the number of subpaths generated from each path that reaches A_{i-1}. The splitting estimator of γ is

$$\hat{\gamma} = \hat{\gamma}(R_1, \cdots, R_k) = \frac{1}{R_1 \cdots R_k} \sum_{i_1=1}^{R_1} \cdots \sum_{i_k=1}^{R_k} I(i_1, \cdots, i_k) \triangleq \frac{Z_k}{R_1 \cdots R_k} \tag{1}$$

where $I(i_1, \cdots, i_k)$ is defined recursively as follows.

1) $I(i_1) = 1$ if the i_1th path from level 0 hits level 1; it is 0 otherwise. If $I(i_1) = 1$, then R_2 subpaths are started and the i_2th of these is labeled (i_1, i_2).
2) If $I(i_1, \cdots, i_{j-1}) = 0$ then $I(i_1, \cdots, i_{j-1}, i_j) = 0$. If $I(i_1, \cdots, i_{j-1}) = 1$ and path $(i_1, \cdots, i_{j-1}, i_j)$ hits level j, then $I(i_1, \cdots, i_{j-1}, i_j) = 1$; it is 0 otherwise. If path $(i_1, \cdots, i_{j-1}, i_j)$ hits level j then R_{j+1} subpaths are started and the i_{j+1}th of these is labeled $(i_1, \cdots, i_{j-1}, i_j, i_{j+1})$.

Clearly, each $I(i_1, \ldots, i_k)$ has expectation $p_1 \cdots p_k = \gamma$. It follows that $\hat{\gamma}$ has the same expectation and is therefore unbiased.

II. Asymptotic Efficiency: Small Interlevel Upcrossing Probabilities

In this section, we consider splitting with a fixed number of intermediate levels, specifically two levels. Reaching the set $A = A_k$ becomes rare because the probability of reaching each level from the previous one goes to zero with a rarity parameter $\epsilon \to 0$. This setting is typical of models of highly dependable fault-tolerant computing systems [29], where reaching a failed state becomes increasingly rare as the failure rates of individual components decrease to zero. Settings in which reaching A_k becomes rare because k increases are treated in Section IV; these are more typical of buffer overflow problems and reliability systems with a high degree of redundancy.

Our first objective is to formalize the intuitive discussion of Fig. 2 in Section I, beginning with the case $k = 2$ illustrated there. We show that if there is a state b such that entering level 1 in state b is highly unlikely, but entering it via b becomes likely conditional on level 2 being reached, then a loss in efficiency occurs. This loss in efficiency comes about because only rarely does the method concentrate its effort on trials from an important state (b) at level 1.

As described in Section I-B, the estimate $\hat{\gamma} = \hat{\gamma}_2$ is obtained by simulating R_1 samples from A_0 and R_2 splits for each success (or *hit*) at level 1. Specifically, the splitting estimator for the two-level case is

$$\hat{\gamma}(R_1, R_2) = \frac{1}{R_1 R_2} \sum_{i_1=1}^{R_1} \sum_{i_2=1}^{R_2} I(i_1, i_2). \tag{2}$$

As described in [14], the two-level splitting procedure can be viewed as picking a *random* probability $\tilde{p}_2$ for success in going from level 1 to 2, where $E[\tilde{p}_2] = p_2$. Conditional on the entry state at level 1, the second-stage hit probability is constant, but because the entry state is unknown at the start, the situation is equivalent to choosing a random $\tilde{p}_2$. All R_2 subpaths generated from a single hit at level 1 have the same probability of reaching level 2, determined by their common entrance state at level 1.

To study the relative efficiency of the method we are interested in the behavior of $\eta = w\sigma^2/\gamma^2$ where w is the expected work to obtain the estimate and σ^2 denotes the variance of $\hat{\gamma}$ (see [18], [20], and Section IV-A for a further discussion of this performance measure). In our case, assuming a constant computational cost $(=1)$ for each split path, we have $w = R_1(1 + p_1 R_2)$. Using an expression for σ^2 derived in [14] we obtain

$$\eta = \left[\frac{1 + p_1 R_2}{p_1}\right]\left[(1 - p_1) + \frac{1 - p_2}{p_2 R_2} + \frac{\mathrm{Var}(\tilde{p}_2)(R_2 - 1)}{p_2^2 R_2}\right]. \tag{3}$$

We are interested in the behavior of η when the p_i's tend to zero, which as explained above might be the case in simulations of highly dependable fault-tolerant computer system models [29]. We index the probabilities by ϵ and assume that as $\epsilon \to 0$, the transition probabilities change to make the set A increasingly rare. Since A is fixed, for the asymptotics, we can also consider the k and the A_i's, $1 \leq i \leq k-1$, to be fixed as ϵ varies. Assume that $p_i = \Omega(\epsilon^{r_i})$ where the r_i's govern the

relative rates with which the p_i's approach zero. We see that (3) grows at a rate of at least $1/p_1$. This lower bound on the growth rate is achieved provided $\mathrm{Var}(\tilde{p}_2)/p_2^2 = O(1)$, $p_1R_2 = O(1)$ and $1/(p_2R_2) = O(1)$. These last two conditions hold if, e.g., $R_2 = 1/p_2$ and $p_1 \le cp_2$ for some constant c. Thus, the splitting procedure does not satisfy the bounded relative error property [29], which would require (3) to remain bounded as $\epsilon \to 0$. However, for standard simulation ($R_1 = R_2 = 1$), (3) grows at rate $1/p_1p_2$, so at best, two-level splitting can be of order $1/p_2$ times more efficient than standard simulation.

To understand the circumstances under which the lowest possible growth rate of order $1/p_1$ is obtained for η, we must examine the behavior of $E[\tilde{p}_2^2]/p_2^2$. To do so, we introduce some additional notation. Let

$$\begin{aligned} e_i(a) &= \text{probability that level } i \text{ is entered at state } a \\ e_i(B) &= \sum_{a\in B} e_i(a), \text{for any } B \subset A_i \\ &\quad \text{(in particular, } e_i(A_i) = \gamma_i) \\ \hat{e}_i(a) &= e_i(a)/\gamma_i \\ &= \text{conditional probability of entering level} \\ &\quad i \text{ at state } a, \text{ given level } i \text{ is reached} \\ p_i(a) &= \text{probability of reaching level } i, \\ &\quad \text{given that level } (i-1) \text{ was entered at } a. \end{aligned}$$

With this notation the probability that $\tilde{p}_2 = p_2(a)$ is given by $\hat{e}_1(a)$ and $p_i = \sum_a \hat{e}_{i-1}(a)p_i(a)$, where for simplicity we assume that the process starts out in state $a \in A_0$ with probability $\hat{e}_0(a)$. We thus obtain

$$\frac{E[\tilde{p}_2^2]}{p_2^2} = \frac{\sum_a \hat{e}_1(a)p_2(a)^2}{p_2^2}. \tag{4}$$

Since $p_2(a)^2 \le p_2(a)$, (4) is at most $1/p_2$ which implies that two-level splitting is no worse than standard simulation provided $p_2 = 1/R_2$ and $p_1R_2 = O(1)$. However, determining verifiable conditions under which $E[\tilde{p}_2^2]/p_2^2$ remains bounded is not always a simple matter. One case in which it is true is if $p_2(a) \le cp_2$ for all states a and some constant c; in this case $\tilde{p}_2/p_2$ is bounded. The following theorem provides additional insight into more general situations. For any state $a \in A_i$, let

$$\begin{aligned} \hat{c}_{i,k}(a) &= \text{probability of entering level } i \text{ at } a \\ &\quad \text{conditional on eventually reaching level } k \\ \hat{c}_{i,k}(B) &= \sum_{a\in B} \hat{c}_{i,k}(a), \text{for any } B \subset A_i. \end{aligned}$$

Hence $\hat{c}_{1,2}(a) = e_1(a)p_2(a)/(p_1p_2)$. Finally, let $\bar{p}_2(B) = \sup\{p_2(b) : b \in B\}$. We now have the following.

Theorem 1:

1) If for some state $b \in A_i$, we have $\hat{c}_{1,2}(b)p_2(b)/p_2 \to \infty$, then $E[\tilde{p}_2^2]/p_2^2 \to \infty$ as $\epsilon \to 0$.
2) Suppose there exists a state $b \in A_i$ such that, as $\epsilon \to 0$

$$\hat{e}_1(b) \to 0 \quad \text{and} \quad \hat{c}_{1,2}(b) \to \alpha > 0. \tag{5}$$

Then $E[\tilde{p}_2^2]/p_2^2 \to \infty$ as $\epsilon \to 0$.

3) Suppose there is a finite set G such that $\liminf_{\epsilon\to 0} \hat{e}_1(a) > 0$ for all $a \in G$. Then $E[\tilde{p}_2^2]/p_2^2 = O(1)$ provided

$$\hat{c}_{1,2}(G^c)\frac{\bar{p}_2(G^c)}{p_2} = O(1). \tag{6}$$

Before proving the result, we interpret its meaning. Let us say that a state b is "on a likely path to A_2" if $\hat{c}_{1,2}(b)$ remains bounded away from zero as $\epsilon \to 0$. Part 1 indicates that $p_2(b)/p_2$ must not be too large, relative to the conditional probability of hitting level 2 by passing through state b at level 1. In particular, if b is on a likely path to A_2, then $p_2(b)/p_2$ must remain bounded, i.e., for any such state b the probability of going from b to level 2 must not be much larger than the average probability of going to level 2. Even if b is not on a likely path to A_2, there are limits on how large $p_2(b)/p_2$ may be. Part 2 states that if entering level 1 at state b is unlikely, but b is on a likely path to A_2, then the method loses efficiency, supporting the discussion around Fig. 2 in Section I. Consider Part 3 with a finite set G such that $\hat{e}_1(G) \to 1$ and suppose further that G^c is not on a likely path to A_2, i.e., $\hat{c}_{1,2}(G^c) \to 0$. Thus passing through the set G^c contributes insignificantly to γ_2 and the simulation spends a negligible fraction of its time simulating splits from G^c (since $\hat{e}_1(G^c) \to 0$). However, even in this case it is not guaranteed that $E[\tilde{p}_2^2]/p_2^2 < \infty$. Part 3 gives a sufficient condition, in terms of the worst case $p_2(b)/p_2$.

Proof: The proof of Part 1 is immediate since each term in the summation of (4) must be finite. To prove Part 2, we have

$$\begin{aligned} E[\tilde{p}_2^2]/p_2^2 &\ge \frac{\hat{e}_1(b)p_2(b)^2}{p_2^2} = \frac{e_1(b)p_2(b)^2}{p_1p_2^2} \\ &= \frac{e_1(b)^2p_2(b)^2}{p_1^2p_2^2}\frac{p_1}{e_1(b)} = (\hat{c}_{1,2}(b))^2\frac{p_1}{e_1(b)} \end{aligned} \tag{7}$$

which $\to \infty$ by (5). To prove Part 3, write (4) as $S_G + S_{G^c}$ where S_G (respectively, S_{G^c}) is the sum over terms in G (respectively, G^c). Since $p_2 \ge \hat{e}_1(a)p_2(a)$, for any $a \in G$ we have $p_2(a)/p_2 \le 1/\hat{e}_1(a)$, and thus S_G is $O(1)$ by the definition of G and the fact that G is finite. Next

$$\begin{aligned} S_{G^c} &= \frac{\sum_{b\in G^c} \hat{e}_1(b)p_2(b)^2}{p_2^2} \le \frac{\sum_{b\in G^c} e_1(b)p_2(b)}{p_1p_2}\frac{\bar{p}_2(G^c)}{p_2} \\ &= \hat{c}_{1,2}(G^c)\frac{\bar{p}_2(G^c)}{p_2}. \end{aligned} \tag{8}$$

□

III. Apparent Bias in Splitting Estimates

As noted in Section I-B, the splitting estimator $\hat{\gamma}$ in (1) is an unbiased estimator of γ, i.e., $E[\hat{\gamma}] = \gamma$. However, now we will show that unless the levels are chosen consistent with the most likely path, the estimator appears biased with high probability, even for large sample sizes.

We consider a family of problems indexed by a rarity parameter ϵ. Associated with each ϵ are splitting factors $R_j(\epsilon)$, an intermediate level $L(\epsilon)$, a final level $k(\epsilon)$, and a particular subset, $\tilde{A}_\epsilon$, of $A_{L(\epsilon)}$. For notational simplicity, we will typically suppress the dependency of R_j, L, k, and $\tilde{A}$ on

ϵ. R_1 is best thought of as the total number of replications of the splitting procedure starting at level zero, i.e., we view the procedure as consisting of R_1 i.i.d. replications where each replication consists of following all the offspring from a single path starting at level 0. As $\epsilon \to 0$, we seek a situation such as that shown in Fig. 2 with the set $\tilde{A}$ representing a neighborhood about the point b: entrance to the intermediate level in $\tilde{A}$ is rare, but entrance to the intermediate level in $\tilde{A}$ given that the final level is hit is not rare. We imagine two ways in which this happens.

1) L and k remain fixed, but the probability of hitting the next level goes to 0 as $\epsilon \to 0$. This is similar to the setting of Section II and is represented by models of highly dependable computing systems.
2) The probability of hitting the next level is independent of ϵ, but L and k increase as $\epsilon \to 0$. This is similar to the multilevel setting of Section IV and is represented by buffer overflow models.

Theorem 2: Suppose

$$e_L(\tilde{A}) \to 0 \quad \text{and} \quad \hat{c}_{L,k}(\tilde{A}) \to \alpha > 0 \quad \text{as } \epsilon \to 0. \tag{9}$$

If $R_1 \to \infty$ and $R_1 \cdots R_L e_L(\tilde{A}) \to 0$ as $\epsilon \to 0$, then there exists a set F_ϵ such that

$$\lim_{\epsilon \to 0} P(F_\epsilon) = 1 \quad \text{and} \quad \limsup_{\epsilon \to 0} \frac{E[\hat{\gamma}\mathbf{1}_{F_\epsilon}]}{\gamma} \le 1 - \alpha \quad \text{as } \epsilon \to 0.$$

Remarks: Condition (9) states that entering A_L via $\tilde{A}$ is unlikely, whereas, conditional on reaching A_k, entering A_L in $\tilde{A}$ becomes likely. The expression $R_1 \cdots R_L e_L(\tilde{A})$ is simply the expected number of subpaths in the first R_1 replications that enter $\tilde{A}$; we assume that R_1 is large, but small enough so that entrances to $\tilde{A}$ are still rare. The theorem states that under the given conditions, with probability approaching one, $\hat{\gamma}$ appears biased. More specifically, while $\hat{\gamma}$ is unbiased, with high probability the event F_ϵ (as defined below) occurs. Thus with high probability, the estimate of γ produced by the simulation is $\hat{\gamma}\mathbf{1}_{F_\epsilon}$. Because the intermediate level is chosen incorrectly, $E[\hat{\gamma}\mathbf{1}_{F_\epsilon}] < \gamma$, i.e., the estimate "appears biased." If $\alpha = 1$, then the process does not pass through the "correct" intermediate set of states $\tilde{A}$ on its way to A_k with probability approaching one. In this case, the splitting estimator appears to arbitrarily underestimate γ. Note also, that the result remains true if in (9) the unconditional probability $e_L(\tilde{A}) \to 0$ is replaced by the conditional probability $\hat{e}_L(\tilde{A}) \to 0$ [since $e_L(\tilde{A}) \le \hat{e}_L(\tilde{A})$].

Proof: Let τ_ϵ be the index of the first replication from zero such that at least one subpath hits $\tilde{A}$ and define $F_\epsilon = \{\tau_\epsilon \ge R_1\}$. First we will show that $P(F_\epsilon) \to 1$ as $\epsilon \to 0$. Note that τ_ϵ has a geometric distribution with some success probability, say $\tilde{\delta}$. Let $\tilde{Z}$ be the total number of offspring from level 0 on a replication that enter level L in $\tilde{A}$. Note that $E[\tilde{Z}] = R_2 \cdots R_L e_L(\tilde{A})$. Furthermore, $\tilde{\delta} = P(\tilde{Z} \ge 1) \le E[\tilde{Z}]$. Therefore $P(\tau_\epsilon > j) = (1 - \tilde{\delta})^j \ge (1 - E[\tilde{Z}])^j$. Hence

$$P(F_\epsilon) \ge (1 - E[\tilde{Z}])^{R_1} = \left(1 - \frac{R_1 \cdots R_L e_L(\tilde{A})}{R_1}\right)^{R_1} \to \exp(0) = 1$$

as $\epsilon \to 0$ since, by hypothesis, $R_1 \to \infty$ and $R_1 \cdots R_L e_L(\tilde{A}) \to 0$.

For the next part, define $Z_k(i)$ to be the total number of offspring from level 0 on replication i that enter level k and let $Y_k(i)$ denote the number of these that pass through $\tilde{A}^c$. Note that

$$E[\hat{\gamma}\mathbf{1}_{F_\epsilon}] = \frac{1}{R_1 \cdots R_k} \sum_{i=1}^{R_1} E\big[Z_k(i)\mathbf{1}_{F_\epsilon}\big] = \frac{1}{R_1 \cdots R_k} \sum_{i=1}^{R_1} E\big[Y_k(i)\mathbf{1}_{F_\epsilon}\big] \tag{10}$$

since on F_ϵ, $Z_k(i) = Y_k(i)$, i.e., all subpaths that reach level k must do so by passing through $\tilde{A}^c$ since on F_ϵ no subpath even enters $\tilde{A}$. Furthermore

$$E\big[Y_k(i)\mathbf{1}_{F_\epsilon}\big] \le E[Y_k(i)] = R_2 \cdots R_k \sum_{a \in \tilde{A}^c} e_L(a) p_{L,k}(a) \tag{11}$$

where $p_{L,k}(a)$ is the probability of reaching level k after entering level L in state a. Thus dividing (10) by γ, using (11), and letting $\epsilon \to 0$, we obtain

$$\frac{E\big[\hat{\gamma}\mathbf{1}_{F_\epsilon}\big]}{\gamma} \le \frac{\sum_{a \in \tilde{A}^c} e_L(a) p_{L,k}(a)}{\gamma} = \hat{c}_{L,k}(\tilde{A}^c) \to 1 - \alpha$$

by (9). □

IV. Asymptotic Efficiency: Large Number of Levels

In Section II we considered the splitting method applied to systems where the rare set was fixed and the probability of moving from one level to the next is small. As already noted, models of highly reliable computing systems provide motivation for this case. In such settings, insight into the performance of the method is obtained by keeping the intermediate sets fixed and considering asymptotics as the intermediate level hitting probabilities tend to zero. In this section, we consider cases in which the probability of moving from one level to the next is not very small, but the number of levels is potentially large. Overflow events in queues with large buffers provide motivation for this case. We now consider asymptotics in which we keep the A_i's fixed but let k increase to infinity—$1/k$, rather than ϵ, is now the rarity parameter.

We derive conditions that must be satisfied in order for multilevel splitting to be asymptotically optimal as $k \to \infty$. We begin with a formal definition of asymptotic optimality and then derive our first necessary condition. This condition specifies a unique value for the number R of splits per level (assumed constant across levels). However, even if this value of R is selected, the method is not guaranteed to be asymptotically optimal. We develop necessary conditions on the large deviations rate functions to enter an intermediate level near some point, $\delta_1(x)$, and the rate, $\delta_2(x)$, to move from that point to the final level. This necessary condition requires that, for all x, $\delta_2(x)$ cannot be too large relative to $\delta_1(x)$ and the overall probability of the rare event being estimated.

In particular, this implies the analog of Part 2 of Theorem 1: the method is not asymptotically optimal if there is a point x at

some intermediate level such that the process is highly unlikely to enter the intermediate level near x, given the *intermediate* level is hit, but it is likely that the process passes near x, given the *final* level is hit. In this sense, the intermediate levels should be chosen so as to be consistent with the large deviations path (assuming it exits) to the *final* level, i.e., if the process is likely to pass near x conditional on reaching the final level, then it must also be likely to enter the intermediate level near x conditional on reaching the intermediate level. We then give an example of a process and level structure such that the necessary conditions are satisfied.

A. Asymptotic Optimality

We will call a sequence of estimators $\{\hat{\gamma}_k\}_{k=1}^{\infty}$ *asymptotically optimal* if

$$\lim_{k\to\infty} \frac{\log\big(E[\hat{\gamma}_k^2]w(k)\big)}{\log(\gamma_k)} = 2 \tag{12}$$

with $w(k)$ denoting the expected computational effort per replication of $\hat{\gamma}_k$. If γ_k has an exponential decay rate, (12) requires that the product of the second moment of $\hat{\gamma}_k$ and the expected computational effort per replication have a decay rate twice as large. In balancing estimator variance and computational effort, it is conventional in simulation to consider the *work-normalized variance* $\mathrm{Var}[\hat{\gamma}_k]w(k)$, rather than $E[\hat{\gamma}_k^2]w(k)$. (This criterion dates at least to Hammersley and Handscomb [20] and is justified in a general framework by Glynn and Whitt [18].) For this reason, in [13] we used the condition

$$\lim_{k\to\infty} \frac{\log(\mathrm{Var}[\hat{\gamma}_k]w(k))}{\log(\gamma_k)} = 2. \tag{13}$$

A simple consequence of Jensen's inequality is

$$\lim_{k\to\infty} \frac{\log(\mathrm{Var}[\hat{\gamma}_k]w(k))}{\log(\gamma_k)} \geq \lim_{k\to\infty} \frac{\log\big(E[\hat{\gamma}_k^2]w(k)\big)}{\log(\gamma_k)} \geq 2 \qquad \forall k \geq 1$$

from which it is evident that (13) implies (12) while the failure of (12) implies that of (13). Since our focus here is on necessary conditions (and in [13] it was on sufficient conditions) it is appropriate to work with the somewhat simpler requirement (12).

We will not make detailed assumptions about the computational effort $w(k)$. Instead, we assume (rather conservatively) that each path started at any level consumes at least one unit of computing time, so that $w(k)$ grows at least as fast as the number of paths.

B. Necessary Conditions on the Number of Splits

In this section, we derive a necessary condition on the number of splits per level in order for multilevel splitting to be asymptotically optimal. We assume that γ_k has a logarithmic limit, i.e.,

$$\lim_{k\to\infty} \frac{\log(\gamma_k)}{k} = \log(\rho) \tag{14}$$

for some constant ρ, $0 < \rho < 1$. This holds quite generally; in particular, it holds in the settings treated in [13] and in the Jackson network considered in Section V.

We assume that the splitting factor at each level is R and let Z_k be the number of paths that enter level k. We can express the (unbiased) splitting estimator of γ_k as

$$\hat{\gamma}_k = \frac{Z_k}{R^k}. \tag{15}$$

Taking expectations of (15) and using (14), we have for any value of R

$$\log(\rho) = \lim_{k\to\infty} \frac{\log(E[Z_k])}{k} - \log(R) = \log(\mu) - \log(R) \tag{16}$$

where $\log(\mu) = \lim_{k\to\infty} \log(E[Z_k])/k$. Write $E[Z_k] = P(Z_k > 0)E[Z_k \mid Z_k > 0] = \hat{p}_k E[Z_k \mid Z_k > 0]$ where $\hat{p}_k = P(Z_k > 0)$. We further assume the limits

$$\lim_{k\to\infty} \frac{\log(E[Z_k \mid Z_k > 0])}{k} \triangleq \log(\hat{\mu})$$

and

$$\lim_{k\to\infty} \frac{\log(P(Z_k > 0))}{k} \triangleq \log(\hat{p}) \tag{17}$$

exist, in which case $\log(\mu) = \log(\hat{\mu}) + \log(\hat{p})$. We remark that in case these limits fail to exist, by considering appropriate subsequences the following theorem continues to hold with the definitions of $\hat{\mu}$ and $\hat{p}$ appropriately modified.

Theorem 3: If (14) and (17) hold, then a necessary condition for splitting to be asymptotically optimal is

$$\log(\hat{p}) = \log(\hat{\mu}) = 0$$

in which case $\log(\mu) = 0$ and hence $R = 1/\rho$.

Proof: We will show that splitting cannot be asymptotically optimal if either $\log(\mu) > 0$ or $\log(\mu) < 0$. First, suppose $\log(\mu) > 0$. We need to show that $\liminf_{k\to\infty} \log(w(k)E[\hat{\gamma}_k^2])/k > 2\log(\rho)$ where $w(k)$ is the expected work. Since $\liminf_{k\to\infty} \log(E[\hat{\gamma}_k^2])/k \geq 2\log(\rho)$ (the variance is nonnegative), it suffices to show that $\liminf_{k\to\infty} \log(w(k))/k > 0$. However, this follows immediately since $w(k) \geq E[Z_k]$. Now, suppose $\log(\mu) < 0$, in which case the expected number of paths to reach level k is exponentially small. There are two cases to consider: $\log(\hat{\mu}) > 0$ and $\log(\hat{\mu}) = 0$. We first consider the case when $\log(\hat{\mu}) > 0$. (In this case, $Z_k = 0$ with high probability, but when $Z_k > 0$, Z_k can be very large.) We will show that $\liminf_{k\to\infty} \log(E[\hat{\gamma}_k^2])/k > 2\log(\rho)$. The unbiasedness of $\hat{\gamma}_k$ implies that

$$\log(\rho) = \log(\hat{p}) + \log(\hat{\mu}) - \log(R). \tag{18}$$

Since $E[Z_k^2] = P(Z_k > 0)E[Z_k^2 \mid Z_k > 0]$ and $E[Z_k^2 \mid Z_k > 0] \geq E[Z_k \mid Z_k > 0]^2$, we have

$$\begin{aligned}
&\liminf_{k\to\infty} \frac{\log\big(E[\hat{\gamma}_k^2]\big)}{k} \\
&\quad = \liminf_{k\to\infty} \frac{\log\big(E[Z_k^2]\big)}{k} - 2\log(R) && (19)\\
&\quad \geq \log(\hat{p}) + 2\log(\hat{\mu}) - 2\log(R) && (20)\\
&\quad > 2(\log(\hat{p}) + \log(\hat{\mu}) - \log(R)) = 2\log(\rho) && (21)
\end{aligned}$$

where the strict inequality follows since $0 > \log(\hat{p})$ and the last equality is true by (18).

We next consider the case $\log(\hat{\mu}) = 0$, in which case $\log(\mu) = \log(\hat{p})$. Let $Y_k = \min(Z_k, 1)$; Y_k is Bernoulli with success probability $\hat{p}_k$. Thus $E[Z_k^2] \geq E[Y_k^2] = \hat{p}_k$ and

$$\begin{aligned}\liminf_{k\to\infty} \frac{\log\left(E\left[\hat{\gamma}_k^2\right]\right)}{k} &\geq \log(\hat{p}) - 2\log(R) = \log(\mu) - 2\log(R) \quad (22)\\ &> 2(\log(\mu) - \log(R)) = 2\log(\rho). \quad (23)\end{aligned}$$

Note that the line of reasoning in (19)–(21) also shows that splitting is not asymptotically optimal if $\log(\mu) = 0$, but $\log(\hat{\mu}) > 0$ and $\log(\hat{p}) < 0$.

Theorem 3 gives a necessary condition for splitting to be asymptotically optimal. In the Markovian cases considered in [13], the necessary condition $R = 1/\rho$ is also sufficient for asymptotic optimality. If $R = 1/\rho$, the requirement that $\log(\hat{p}) = 0$ means that the probability of getting at least one success at level k cannot be too small while the requirement that $\log(\hat{\mu}) = 0$ means that, given at least one success at level k, the expected number of successes must be subexponential. Roughly speaking, if $R < 1/\rho$ then there are not enough splits so that entering the final level is still a rare event, thereby precluding asymptotic optimality. If $R > 1/\rho$, then the expected number of subpaths entering the final level grows exponentially. This exponential increase in the work precludes asymptotic optimality. In the finite Markovian case, if $R = 1/\rho$, then the expected number of splits entering each level remains roughly constant. In this more general setting, $R = 1/\rho$ implies that the expected number of splits entering each level neither grows nor shrinks too quickly.

Note also that the theorem remains valid if, instead of a constant R splits per level, there are R_j splits at level j and $\log(R_1 \cdots R_k)/k \to \log(R)$. Furthermore, results in [13] imply that if the number of splits at level j is random, i.i.d. with mean $E[R_j]$ and independent of everything else, then $\hat{\gamma}_k = Z_k/(E[R_1] \cdots E[R_k])$ is an unbiased estimate of γ_k. Theorem 3 thus remains valid if $\log(E[R_1] \cdots E[R_k])/k \to \log(R)$, i.e., a necessary condition for asymptotic optimality is still $R = 1/\rho$.

C. Necessary Conditions on the Rate Functions for Asymptotic Optimality

In this subsection we consider necessary conditions on the rate functions for entering an intermediate level near a point and moving from that point to the final level (see, e.g., Dembo and Zeitouni [9] for background on rate functions). In proving our results we will rely on the following simple lemma.

Lemma 1: If (14) and (17) hold, a necessary condition for asymptotic optimality is that

$$\lim_{k\to\infty} \frac{\log\left(E\left[Z_k^2\right]\right)}{k} = 0.$$

Proof: By Theorem 3 it follows that we must have $\log(R) = -\log(\rho)$. We then have, using that $w(k) \geq 1$, $\hat{\gamma}_k$ is unbiased, and Jensen's inequality, that

$$\begin{aligned}2\log(\rho) &\leq \frac{\log\left(E\left[\hat{\gamma}_k^2\right]\right)}{k} = \frac{\log\left(E\left[Z_k^2\right]\right)}{k} - 2\log(R)\\ &= \frac{\log\left(E\left[Z_k^2\right]\right)}{k} + 2\log(\rho).\end{aligned}$$

From this the statement of the lemma follows. □

The setting for the remainder of this subsection is as follows. A fixed set A is scaled by a parameter k, and the resulting set kA becomes rare as k increases. In simulating the probability of hitting the set kA, a set B has been chosen so that $[\alpha k]B \subset kA$ and the $[\alpha k]$th splitting occurs upon entry into $[\alpha k]B$. (For notational simplicity, we henceforth assume that $k \to \infty$ in such a way that αk is an integer, although the result is true in general provided $\alpha k B$ is used to denote $[\alpha k]B$.) A special case is when $\alpha = 1/2$ in which case we are considering the $k/2$th splitting to take place at $(k/2)B$.

The assumptions of the following theorem state that there exists an x such that the probability of entering the set $\alpha k B$ in an appropriately chosen neighborhood of $\alpha k x$ satisfies a large deviations lower bound. In addition, the probability of hitting the set A starting from any of the points in this neighborhood satisfies a uniform large deviations lower bound. The result, (26), places a constraint on the relative magnitudes of these lower bounds. For example, if entering $\alpha k B$ in a neighborhood of $\alpha k x$ is unlikely, it cannot be too "easy" to then enter kA from this neighborhood.

We let $\{kA\}$ denote the event that kA has been entered and $\{\alpha k B; y\}$ and $\{\alpha k B; B(\alpha k x, k\epsilon)\}$ denote the events that $\alpha k B$ has been entered at y and at $B(\alpha k x, k\epsilon) \triangleq \{z : |z - \alpha k x| \leq k\epsilon\}$, respectively. □

Theorem 4: Fix x and suppose that for some positive $\delta_1(x)$ and $\delta_2(x)$ we have

$$\lim_{\epsilon\to 0} \liminf_{k\to\infty} \frac{1}{k} \log(P(\{\alpha k B; B(\alpha k x, k\epsilon)\})) \geq \alpha \log(\delta_1(x)) \quad (24)$$

and

$$\begin{aligned}\lim_{\epsilon\to 0} \liminf_{k\to\infty} \inf_{B(\alpha k x, k\epsilon)} \frac{1}{k} \log(P(\{kA\} \mid \{\alpha k B; y\}))\\ \geq (1-\alpha)\log(\delta_2(x)). \quad (25)\end{aligned}$$

Then, for asymptotic optimality to hold it is necessary that

$$\alpha \log(\delta_1(x)) + 2(1-\alpha)\log(\delta_2(x)) \leq (2-\alpha)\log(\rho). \quad (26)$$

Proof: Let $N_1(y)$ be the number of paths that enter $\alpha k B$ at y and let $Z_i(y)$, $i = 1, \ldots, N_1(y)$ be the number of successors of the ith of these that reaches kA. For every $\epsilon > 0$ we have the relation

$$Z_k \geq \sum_{y \in B(\alpha k x, k\epsilon)} \left(\sum_{i=1}^{N_1(y)} Z_i(y) \right)$$

and the $Z_i(y)$ are i.i.d. This implies

$$Z_k^2 n \geq \sum_{y \in B(\alpha k x, k\epsilon)} \sum_{i=1}^{N_1(y)} Z_i(y)^2$$

and Wald's equation then allows us to write

$$EZ_k^2 \geq \sum_{y \in B(\alpha k x, k\epsilon)} E[N_1(y)]E[Z(y)]^2$$

where $Z(y)$ has the distribution of the $Z_i(y)$. Since $N_1(y)/R^{\alpha k}$ is an unbiased estimator of $P(\{\alpha kB; y\})$

$$\frac{E[N_1(y)]}{R^{\alpha k}} = P(\{\alpha kB; y\}).$$

Similarly

$$\frac{E[Z(y)]}{R^{(1-\alpha)k}} = P(\{kA\} \mid \{\alpha kB; y\}).$$

Therefore, we may write

$$\begin{aligned}\lim_{k\to\infty} \frac{1}{k} \log\bigl(EZ_k^2\bigr) &\geq (2-\alpha)\log R + 2 \liminf_{k\to\infty} \inf_{B(\alpha k x, k\epsilon)} \\ &\quad \times \frac{1}{k}\log(P(\{kA\} \mid \{\alpha kB; y\})) \\ &\quad + \liminf_{k\to\infty} \frac{1}{k}\log(P(\{\alpha kB; B(\alpha k x, k\epsilon)\})).\end{aligned}$$

Letting $\epsilon \to 0$ and recalling Theorem 3 and Lemma 1 yields the desired result. □

Note that we always have

$$\alpha \log(\delta_1(x)) + (1-\alpha)\log(\delta_2(x)) \leq \log(\rho) \tag{27}$$

since the probability of entering level k by passing near a particular point x is less than the overall probability of entering level k. Comparing (26) and (27), we see that (26) places additional restrictions. In particular, the difference between (26) and (27) is that the left-hand side (LHS) of (26) contains an additional factor of $(1-\alpha)\log(\delta_2(x))$, while the right-hand side (RHS) of (26) contains an additional factor of $(1-\alpha)\log(\rho)$. Thus (26) may fail to hold if $\delta_2(x) \gg \rho$, i.e., if it is too "easy" to reach the final level from x.

In the following theorem the assumption is made that a point x exists such that it is likely to hit the final level by passing near the point x at the intermediate level. (Technically (28) states that the probability that αkB is hit in $B(\alpha k x, k\epsilon)$, given that kA is hit, decays subexponentially and the large deviations lower bounds of Theorem 4 are supplemented with corresponding upper bounds of (29) and (30).) In this case the upper bound of (27) becomes an equality (31). More importantly, we may then rephrase the necessary condition of Theorem 4 as (32), which is best understood if $B = A$, in which case the intermediate set is a scaled version of the final set. If this is the case, the probability of hitting αkB has rate $\alpha\log(\rho)$ and the probability of hitting αkB near x has rate $\alpha\log(\delta_1(x))$. The necessary condition (32) then states that if $\log(\delta_1(x)) < \log(\rho)$, then the method cannot be asymptotically optimal. In other words, the sort of situation illustrated in Fig. 2 must be precluded for all intermediate levels. This is the analog of Part 2 of Theorem 1, since it states that the method is not asymptotically optimal if the most likely point to enter an intermediate level is not contained in the most likely path to the final level.

Theorem 5: Suppose there exists x such that in addition to (24) and (25) it holds that for every $\epsilon > 0$

$$\lim_{k\to\infty} \frac{1}{k}\log\left(\sum_{y\in B(\alpha k x, k\epsilon)} P(\{\alpha kB; y\} \mid \{kA\})\right) = 0 \tag{28}$$

$$\lim_{\epsilon\to 0} \limsup_{k\to\infty} \frac{1}{k}\log(P(\{\alpha kB; B(\alpha k x, k\epsilon)\})) \leq \alpha\log(\delta_1(x)) \tag{29}$$

and

$$\begin{aligned}\lim_{\epsilon\to 0} \limsup_{k\to\infty} \sup_{B(\alpha k x, k\epsilon)} \frac{1}{k}\log(P(\{kA\} \mid \{\alpha kB; y\})) \\ \leq (1-\alpha)\log(\delta_2(x)).\end{aligned} \tag{30}$$

It then follows that

$$\log(\rho) = \alpha\log(\delta_1(x)) + (1-\alpha)\log(\delta_2(x)) \tag{31}$$

and a necessary condition for asymptotic optimality is that

$$\log(\delta_2(x)) \leq \log(\rho) \leq \log(\delta_1(x)). \tag{32}$$

Proof: For every $\epsilon > 0$

$$\begin{aligned}&\sum_{y\in B(\alpha k x, k\epsilon)} P(\{\alpha kB; y\} \mid \{kA\}) \\ &\quad = \sum_{y\in B(\alpha k x, k\epsilon)} \frac{P(\{\alpha kB; y\} \cap \{kA\})}{P(\{kA\})}\end{aligned}$$

and so, by (28)

$$\begin{aligned}&\lim_{k\to\infty}\frac{1}{k}\log(P(\{kA\})) \\ &\quad = \lim_{k\to\infty}\frac{1}{k}\log\left(\sum_{y\in B(\alpha k x, k\epsilon)} P(\{kA\} \right. \\ &\qquad \left. \mid \{\alpha kB; y\})P(\{\alpha kB; y\})\right).\end{aligned}$$

From (24), (25), (29), and (30) it follows that

$$\begin{aligned}&\lim_{\epsilon\to 0}\lim_{k\to\infty}\frac{1}{k}\log\left(\sum_{y\in B(\alpha k x, k\epsilon)} P(\{kA\} \right. \\ &\qquad \left. \mid \{\alpha kB; y\})P(\{\alpha kB; y\})\right) \\ &\quad = \alpha\log(\delta_1(x)) + (1-\alpha)\log(\delta_2(x))\end{aligned}$$

and therefore that

$$\log(\rho) = \alpha\log(\delta_1(x)) + (1-\alpha)\log(\delta_2(x)).$$

Relation (32) now follows upon recalling the necessary condition of Theorem 4. □

D. A Setting Where the Necessary Conditions Are Satisfied

In the previous subsection, a number of highly technical assumptions were made on the rate functions, i.e., (24), (25), (28), (29), and (30). In this section we demonstrate a setting in which these conditions are satisfied. We further show that the optimality condition (32) is satisfied in this setting provided the sets A and B are appropriately defined.

Theorem 6: Suppose the underlying process is given by $\{\sum_{i=1}^{n} Y_i\}_{n=1}^{\infty}$, where $\{Y_i\}_{i=1}^{\infty}$ is an i.i.d. sequence of bounded random variables, A and B are open and convex, and, for simplicity, that $\alpha = 1/2$. Under the conditions of [7, Th. 2.1], (24), (25), (28), (29), and (30) hold.

Proof: That an x satisfying (28) exists is a straightforward consequence of [7, (5.2)] and [37, Th. 4].

We next consider (24) and (29). With $\tau > 0$ fixed, let $n(m) = [m/\tau]$, for m any positive integer. We have that, for any $\epsilon > 0$ and $r > 1$ sufficiently close to 1 (note that $rx \in B$), the existence of an $\tilde{\epsilon} > 0$, independent of k, such that

$$\begin{aligned}\liminf_{k\to\infty} \frac{1}{k}\log(P(\{\alpha k B; B(1/2kx, k\epsilon)\})) \\ \geq \liminf_{k\to\infty} \frac{1}{k}\log\left(P\left(\sup_{0\leq m\leq n(k)} |S_m/n(k) \right.\right. \\ \left.\left. - mrx/2n(k)| \leq \tilde{\epsilon}\right)\right). \qquad (33)\end{aligned}$$

By Mogulskii's large deviation theorem (cf. [9, Ch. 9]) it follows that this last quantity is greater than or equal to $-\tau^{-1} I(\tau r x/2)$, where I is the Fenchel–Legendre transform of the moment-generating function of Y_1. As in the proof of the lower bound part of [7, Th. 2.1], given that $\tau > 0$ was arbitrary, it follows that (33) is greater than or equal to $-\sup_{\{\theta:\Lambda(\theta)=0\}}\langle\theta, rx/2\rangle$, which in turn implies, letting $r \to 1$, that (33) is greater than or equal to $-\sup_{\{\theta:\Lambda(\theta)=0\}}\langle\theta, x/2\rangle$. For (29) we first note that

$$\begin{aligned}\limsup_{k\to\infty} \frac{1}{k}\log(P(\{\alpha k B; B(1/2kx, k\epsilon)\})) \\ \leq \limsup_{k\to\infty} \frac{1}{k}\log(P(B(1/2kx, k\epsilon))) \\ = -\inf_{y\in B(1/2x,\epsilon)} \sup_{\{\theta:\Lambda(\theta)=0\}} \langle\theta, y\rangle\end{aligned}$$

the last equality following from [7, Th. 2.1]. It therefore holds that

$$\begin{aligned}\lim_{\epsilon\to 0}\limsup_{k\to\infty} \frac{1}{k}\log(P(\{\alpha k B; B(kx/2, k\epsilon)\})) \\ \leq -\sup_{\{\theta:\Lambda(\theta)=0\}} \langle\theta, x/2\rangle\end{aligned}$$

and we may set

$$\frac{1}{2}\log(\delta_1(x)) = -\sup_{\{\theta:\Lambda(\theta)=0\}} \langle\theta, x/2\rangle.$$

We now consider (25) and (30) being treated in a similar fashion. It is readily verified that for each $\epsilon > 0$ there exists a $\gamma(\epsilon) > 0$ such that $\lim_{\epsilon\to 0}\gamma(\epsilon) = 0$ and for y such that $|y - kx/2| < k\epsilon$

$$P(\{kA\} \mid \{\alpha k B; y\}) \geq P\big(\{k(A^{\gamma(\epsilon)} - x/2)\}\big)$$

where

$$A^{\gamma(\epsilon)} = \{z \in A : |z - w| > \gamma(\epsilon) \text{ for all } w \in \partial A\}.$$

From [7, Th. 2.1] it follows that

$$\begin{aligned}\lim_{k\to\infty} \frac{1}{k}\log\left(P\big(\{k(A^{\gamma(\epsilon)} - x/2)\}\big)\right) \\ = -\inf_{y\in A^{\gamma(\epsilon)}-x/2} \sup_{\{\theta:\Lambda(\theta)=0\}} \langle\theta, y\rangle.\end{aligned}$$

From this we have that (25) holds with

$$\frac{1}{2}\log(\delta_2(x)) = -\inf_{y\in A-x/2} \sup_{\{\theta:\Lambda(\theta)=0\}} \langle\theta, y\rangle.$$

The same value of $\delta_2(x)$ may be shown to hold for (30). □

Corollary 1: In the setting of Theorem 6, if the set A is a $1/2$-plane and $B = A$, the necessary condition (32) is satisfied.

Proof: It follows from [37] that the point x satisfying (28) is in ∂A and such that

$$\log(\rho) = -\inf_{y\in A} \sup_{\{\theta:\Lambda(\theta)=0\}} \langle\theta, y\rangle = -\sup_{\{\theta:\Lambda(\theta)=0\}} \langle\theta, x\rangle.$$

By the proof of Theorem 6 we see that

$$\begin{aligned}\frac{1}{2}\log(\delta_2(x)) &= -\inf_{y\in A-1/2x} \sup_{\{\theta:\Lambda(\theta)=0\}} \langle\theta, y\rangle \\ &= -\sup_{\{\theta:\Lambda(\theta)=0\}} \langle\theta, 1/2x\rangle\end{aligned}$$

the last equality following since $x \in \partial A$ and A is a 1/2-plane. Recalling again the proof of Theorem 6 we see that $\log(\delta_1(x)) = \log(\delta_2(x))$. We now have that (32) holds, indeed with equalities. □

V. Splitting in a Jackson Network

In this section, we show by means of a counterexample that splitting is not always asymptotically optimal, even when the intermediate sets are consistent with the large deviations behavior.

A. Preliminaries

Consider an m queue Jackson network with traffic intensity $\rho_i < 1$ at queue i. Let $x_i \geq 0$ be given constants (for simplicity assume kx_i is an integer). Let $A_k = \{x_i k \leq Q_i\}$ and let $A_k(D) = \{x_i k \leq Q_i \leq x_i k + D_i\}$ for integers $D_i \geq 0$. Let γ_k ($\gamma_k(D)$) be the probability of hitting $A_k(A_k(D))$ and let $\gamma_k(\epsilon)$ denote the probability of hitting the set $B_k(\epsilon) = \{x_i k \leq Q_i \leq (x_i + \epsilon)k\}$ during a cycle. Let $\tilde{\gamma}_k(\epsilon)$ denote the probability that A_k is first hit somewhere in the set $B(\epsilon)$. Define $\log(\rho) = \sum_{i=1}^{m} x_i \log(\rho_i)$. Let $\pi(A)$ denote the steady-state probability of a set A and let $Y(A) = \int_{s=0}^{\alpha_0} I(X(s) \in A) ds$ where α_0 is the first time to return to the state where $Q_i = 0$ for all i. Then, by regenerative process theory

$$\pi(A) = \frac{E_0[Y(A)]}{E_0[\alpha_0]} = \frac{E_0[Y(A) \mid Y(A) > 0]P\{Y(A) > 0\}}{E_0[\alpha_0]} \qquad (34)$$

where E_0 denotes expectation starting when all queues are empty.

Lemma 2: In a Jackson network with $\rho_i < 1$ for all i:

1) $\lim_{k\to\infty} \log(\gamma_k(D))/k = \log(\rho)$;
2) $\lim_{k\to\infty} \log(\gamma_k)/k = \log(\rho)$;
3) $\lim_{k\to\infty} \tilde{\gamma}_k(\epsilon)/\gamma_k = 1$ for any $\epsilon > 0$;
4) $\lim_{k\to\infty} \log(\tilde{\gamma}_k(\epsilon))/k = \log(\rho)$ for any $\epsilon > 0$.

Proof: To prove 1), apply (34) with $A = A_k(D)$. In this case $P\{Y(A) > 0\} = \gamma_k(D)$. Furthermore, there exist constants c_1 and c_2 such that $c_1 \le E_0[Y(A) \mid Y(A) > 0] \le c_2 k$. The lower bound is true since the process must spend at least one transition in A (given that A is hit). The upper bound is true by [2, Corollary 1] which states that $E_{\mathbf{q}}[\alpha_0] \le c\sum_{i=1}^m q_i$ when the initial queue lengths are q_i. The result then follows since $\pi(A) = \prod_{i=1}^m [\rho_i^{x_i k} - \rho_i^{(x_i k + D_i + 1)}]$. To prove 2), let $A = A_k$, use (34), the lower bound $c_1 \le E_0[Y(A) \mid Y(A) > 0]$ and the form of $\pi(A)$ to show that $\limsup_{k\to\infty} \log(\gamma_k)/k \le \log(\rho)$. For the lower bound on γ_k, since $A_k(D) \subset A_k$, $\gamma_k(D) \le \gamma_k$. Thus, from (34) we obtain

$$\begin{aligned}\pi(A_k(D)) &= \frac{E_0[Y(A_k(D)) \mid Y(A_k(D)) > 0]\gamma_k(D)}{E_0[\alpha_0]} \\ &\le \frac{E_0[Y(A_k(D)) \mid Y(A_k(D)) > 0]\gamma_k}{E_0[\alpha_0]} \le \frac{c_2 k \gamma_k}{E_0[\alpha_0]}.\end{aligned} \tag{35}$$

Using the form of $\pi(A_k(D))$ we thus obtain $\liminf_{k\to\infty} \log(\gamma_k)/k \ge \log(\rho)$. To prove 3), consider $[\gamma_k - \tilde{\gamma}_k(\epsilon)]/\gamma_k$ which is the conditional probability of hitting A_k in $B_k^c(\epsilon)$, the complement of $B_k(\epsilon)$. The numerator in this expression is less than $\gamma_k(B_k^c(\epsilon))$, which is simply the probability of hitting $B_k^c(\epsilon)$ during a cycle. Assume now, without loss of generality, that $\rho_1 \ge \rho_i$ for each i. The various bounds used in proving 1) and 2) imply that $\gamma_k(B_k^c(\epsilon)) = O(\rho_1^{\epsilon k} \prod \rho_i^{x_i k})$. Furthermore, (35) implies that $1/\gamma_k \le ck/[\prod \rho_i^{x_i k}(1+O(1))]$. Thus $\tilde{\gamma}_k(B_k^c(\epsilon))/\gamma_k \to 0$ which is equivalent to 3). Part 4 follows directly from Parts 3 and 2. □

We next give an example of a Jackson network for which the sets $A_j = \{Q_i \ge j\}$ $(x = (1,1))$ are consistent with large deviations in the sense that given that the process reaches A_k, it passes close to the point (ak, ak) with high probability for any $0 < a < 1$. (For notational simplicity, we will assume ak is an integer, but the result is true in general.)

Consider a tandem M/M/1 queue with rates $\lambda < \mu_2 < \mu_1$. Time reversal states that if the reversed process starts at (k,k), it empties Q_1 at rate $\mu_2 - \mu_1$ and Q_2 at rate $\lambda - \mu_2$ until one of the boundaries is hit. Thus the slope $= 1$ if $\mu_2 - \mu_1 = \lambda - \mu_2$. In particular, this is true if $\mu_1 = \lambda + 2\delta$ and $\mu_2 = \lambda + \delta$ where δ is chosen so that $\lambda + \mu_1 + \mu_2 = 1$. In this case, this strongly suggests that the large deviations path builds up at slope $+1$ and therefore the sets A_j are consistent with the large deviations path. The below argument makes this rigorous.

Part 2 of Lemma 2 states that $\log(\gamma_k)/k \to \log(\rho)$ where $\rho = \rho_1\rho_2$. Furthermore, Part 4 of Lemma 2 states that the probability of hitting A_{ak} in a neighborhood about the point (ak, ak) during a cycle, $\tilde{\gamma}_{ak}(\epsilon)$, has rate $a\log(\rho)$. Now, starting at a point $y \in A_{ak}(\epsilon)$, apply importance sampling with rates consistent with time reversal, i.e., $\lambda' = \mu_1$, $\mu_1' = \mu_2$, and $\mu_2' = \lambda$. Consider the event $H_k(\epsilon, T)$ such that the process hits A_k by Tk transitions and for kT transitions stays within a tube of width 2ϵ about the straight line implied by these drifts that goes from (a,a) to $(1,1)$. Formally, in $H_k(\epsilon, T)$, after $[kt]$ transitions

$$\left| \frac{Q_i([kt])}{k} - a - \delta t \right| \le 2\epsilon, \qquad \text{for } i = 1, 2. \tag{36}$$

Under this change of measure, the approximate time to hit A_k satisfies $Tk\delta \approx (1-a)k$, so define $S = (1-a)/\delta$ and $S(\epsilon) = S + 3\epsilon/\delta$. Now

$$P(\{A_k\} \mid y) \ge P\{H_k(\epsilon, S(\epsilon))\} = E'[LI(H_k(\epsilon, S(\epsilon)))] \tag{37}$$

where E' denotes the expectation under the change of measure and L is the likelihood ratio. For small enough ϵ, if $H_k(\epsilon, S(\epsilon))$ occurs, the process stays in the interior and has a simple likelihood ratio

$$\begin{aligned}L &= \left(\frac{\lambda}{\lambda'}\right)^A \left(\frac{\mu_1}{\mu_1'}\right)^{D_1} \left(\frac{\mu_2}{\mu_2'}\right)^{D_2} \\ &= \left(\frac{\lambda}{\mu_1}\right)^A \left(\frac{\mu_1}{\mu_2}\right)^{D_1} \left(\frac{\mu_2}{\lambda}\right)^{D_2} \\ &= \lambda^{A-D_2} \mu_1^{D_1 - A} \mu_2^{D_2 - D_1}\end{aligned} \tag{38}$$

where A is the number of arrivals, D_1 is the number of Q_1 departures (arrivals to Q_2), and D_2 is the number of Q_2 departures. This simplifies to

$$L = \rho_1^{A - D_1} \rho_2^{D_1 - D_2}. \tag{39}$$

Now, on this event at time $kS(\epsilon)$, $k(1-3\epsilon) \le Q_i \le k(1+3\epsilon)$ [by (36) and the definition of $S(\epsilon)$]. Since $Q_1 = y_1 + A - D_1$, $Q_2 = y_2 + D_1 - D_2$, and $k(a-\epsilon) \le y_i \le k(a+\epsilon)$, we obtain $(1-a)k - 4\epsilon k \le A - D_1 \le (1-a)k + 4\epsilon k$ and $(1-a)k - 4\epsilon k \le D_1 - D_2 \le (1-a)k + 4\epsilon k$. Applying these inequalities to (39), we obtain

$$(\rho_1\rho_2)^{k[(1-a)+4\epsilon]} \le L \le (\rho_1\rho_2)^{k[(1-a)-4\epsilon]}. \tag{40}$$

Thus

$$\begin{aligned}\log(P(\{A_k\} \mid y))/k \ge [(1-a) + 4\epsilon]\log(\rho) \\ + \log(E'[I(H_k(\epsilon, S(\epsilon))) \mid y])/k.\end{aligned} \tag{41}$$

We now show that

$$\liminf_{k\to\infty} \inf_{y \in A_{ak}(\epsilon)} \frac{1}{k} \log(E'(I(H_k(\epsilon, S(\epsilon))) \mid y)) = 0. \tag{42}$$

Let y_k be any $y \in A_{ak}(\epsilon)$ so that the infimum in (42) is achieved at y_k (note it suffices to consider only $y \in A_{ak}(\epsilon)$ of the form $y_k = n/k$ for some $n \in \mathbb{N}$, of which there are

only finitely many). We may assume without loss of generality that y_k/k converges to a point $y \in A_a(\epsilon)$. We then have, with $\{Y_i\}$, an i.i.d. sequence each element distributed according to the change of measure

$$\begin{aligned} &E'(I(H_k(\epsilon, S(\epsilon))) \mid y_k) \\ &= P\left(\frac{y_k}{k} + \frac{1}{k}\sum_{i=1}^{kS(\epsilon)} Y_i \in A \cap \right. \\ &\qquad \left. \max_{0\le t\le 1}\left|\frac{y_k}{k} + \frac{1}{k}\sum_{i=1}^{kS(\epsilon)t} Y_i - (a,a) - (\delta,\delta)S(\epsilon)t\right| \le 2\epsilon\right) \\ &= P\left(\max_{0\le t\le 1}\left|\frac{y_k}{k} + \frac{1}{k}\sum_{i=1}^{kS(\epsilon)t} Y_i - (a,a)\right.\right. \\ &\qquad \left.\left. - (\delta,\delta)S(\epsilon)t\right| \le 2\epsilon\right) \end{aligned}$$

where the last equality follows by our choice of $S(\epsilon)$. As $EY_1 = (\delta,\delta)$ and $|\frac{y_k}{k} - (a,a)| \le \epsilon$, (42) follows from the functional law of large numbers. Given this, we may write

$$\liminf_{k\to\infty} \log(P(\{A_k\} \mid y_k))/k \ge [(1-a) + 4\epsilon]\log(\rho) \quad (43)$$

and the desired result follows by letting $\epsilon \to 0$.

B. Counterexample

In this section, we show numerically that the necessary condition of Theorem 4 does not hold for the tandem Jackson network even when the intermediate sets are constructed so as to be consistent with large deviations. As in the previous subsection, we consider the network that is consistent with time reversal. The intermediate level corresponds to $A_{k/2}$. In the following, we will refer to a path from a point x to y; by this we mean a path that starts with Q_i/k within ϵ of x_i and ends with Q_i/k within ϵ of y_i. We consider a path that enters $A_{k/2}$ at $x = (b, 0.5)$ for some $b > 0.5$, i.e., enters $A_{k/2}$ such that Q_i/k is within ϵ of x_i. The necessary condition for asymptotic optimality is then

$$\frac{1}{2}\log(\delta_1(x)) + \log(\delta_2(x)) \le \frac{3}{2}\log(\rho). \quad (44)$$

We will show that a lower bound on the LHS of (44) is greater than the RHS. We do so by considering a particular path that enters level $k/2$ at a point $x > 0.5$. To construct a lower bound on $\delta_1(x)$ consider a path that first goes from $(0,0)$ to (a,a); by part 4 of Lemma 2, this has rate $\log(r_1) = a\log(\rho)$. Next consider a straight line path from (a,a) to $(b, 0.5)$ with rates $(\lambda', \mu_1', \mu_2')$ such that $\lambda' + \mu_1' + \mu_2' = 1$. Let $\log(r_2(\lambda', \mu_1', \mu_2', a, b))$ be the rate associated with this path. Since such a path stays in the interior the likelihood ratio for such a path is given by

$$L = \left(\frac{\lambda}{\lambda'}\right)^A \left(\frac{\mu_1}{\mu_1'}\right)^{D_1} \left(\frac{\mu_2}{\mu_2'}\right)^{D_2}. \quad (45)$$

If the process is simulated for n_k transitions, then in an appropriately defined event whose probability approaches one, we have $|A/n_k - \lambda'| \le \epsilon$ and $|D_i/n_k - \mu_i'| \le \epsilon$. As in the previous section, the approximate number of transitions n_k to reach $(b, .5)$ can be determined; we have $n_k(\lambda' - \mu_1') \approx k(b-a)$ and $n_k(\mu_1' - \mu_2') \approx k(0.5 - a)$ which leads to a formal constraint on the rates

$$\lim_{k\to\infty} \frac{n_k}{k} = \frac{b-a}{\lambda' - \mu_1'} = \frac{0.5 - a}{\mu_1' - \mu_2'}. \quad (46)$$

Combining the above facts shows that

$$\log(r_2(\lambda', \mu_1', \mu_2', a, b)) = -\frac{b-a}{\lambda' - \mu_1'} I(\lambda', \mu_1', \mu_2') \quad (47)$$

is a lower bound on the rate to follow the straight line path where $I(\lambda', \mu_1', \mu_2') = \lambda'\log(\lambda'/\lambda) + \mu_1'\log(\mu_1'/\mu_1) + \mu_2'\log(\mu_2'/\mu_2)$. We can maximize (47), subject to the constraint (46), to obtain the greatest lower bound. Such maximization becomes algebraically complex; however, it can be carried out numerically. In doing so, we observed that the best rate seemed to occur when $\mu_2' = \lambda$ and certainly evaluating r_2 when $\mu_2' = \lambda$ is a lower bound regardless of what the optimum is. Thus fixing $\mu_2' = \lambda$ together with the constraints (46) and $\lambda' + \mu_1' + \mu_2' = 1$ determines λ' and μ_1'; i.e., a lower bound on the optimum rate of $\log(r_2(\lambda', \mu_1', \mu_2', a, b))$ can be written as a function $\log(\underline{r}_2(\lambda, a, b))$. We note that this approach can also be justified by an appeal to Sanov's theorem [3] and is essentially the same heuristic technique presented in [24]. However, it is clear by the analysis of the previous subsection that one can carefully construct the appropriate tubes and make everything perfectly rigorous. Thus we obtain the lower bound $(1/2)\log(\delta_1(x)) \ge a\log(\rho) + \log(\underline{r}_2(\lambda, a, b))$. A similar approach can be taken to obtain a lower bound $(1/2)\log(\delta_2(x)) \ge \log(\underline{r}_3(\lambda, b))$ where $\underline{r}_3$ is again obtained by fixing $\mu_2' = \lambda$ as in (47), but starting at the point $(b, .5)$ and ending at $(1,1)$.

Now define the function $H(a, b, \lambda)$ by

$$\begin{aligned} H(a,b,\lambda) = \frac{3}{2}\log(\rho) - (a\log(\rho) + \log(\underline{r}_2(\lambda, a, b)) \\ + 2\log(\underline{r}_3(\lambda, b))). \end{aligned} \quad (48)$$

(Because of the way the model is parameterized, ρ is a function of λ.) The necessary condition (44) for splitting to be asymptotically optimal cannot be satisfied if $H(a, b, \lambda) < 0$. Fig. 3 plots $H(a, b, \lambda)$ as a function of the arrival rate λ for two fixed pairs of (a, b) : $(0.1, 0.6)$ and $(0.1, 0.7)$. (The X axis in the figure is actually the physically more meaningful traffic intensity at queue 2, which turns out to be $\rho_2 = 3\lambda$.) Fig. 3 shows that $H(a, b, \lambda) < 0$ for all traffic intensities and these values of (a, b). Thus we have shown numerically that splitting cannot be asymptotically optimal for this example.

C. Experimental Results

In this section we consider simulation results when applying splitting to a two-queue tandem Jackson network. The rare event of interest is hitting $A_k = \{Q_1 \ge k, Q_2 \ge k\}$ during

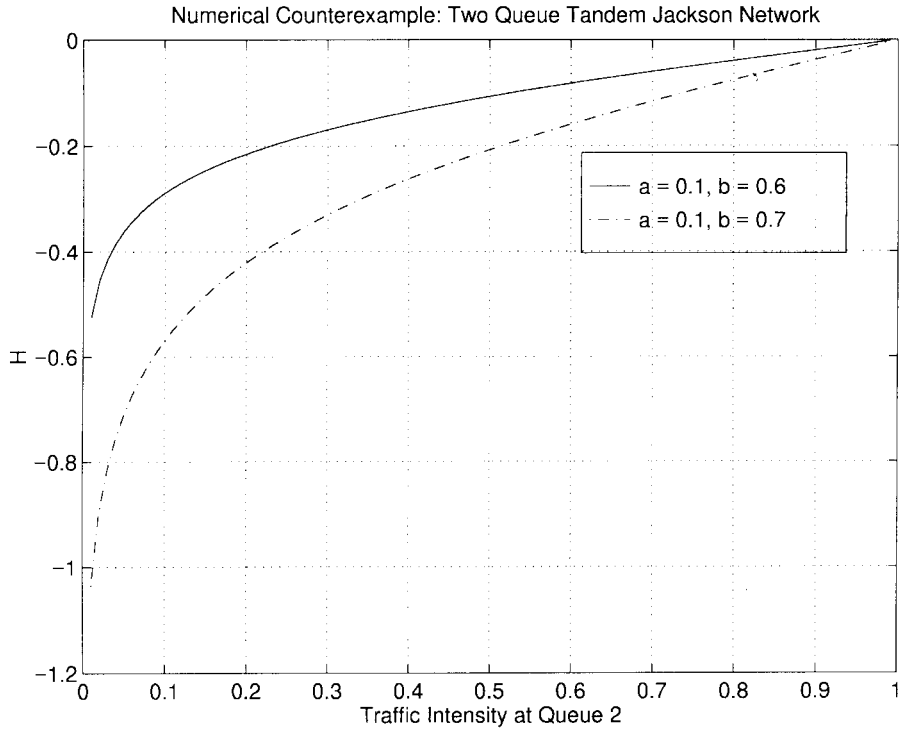


Fig. 3. A numerical counterexample to asymptotic optimality in the two-queue tandem Jackson network showing that $H(a,b,\lambda) < 0$ for all arrival rates and certain values of a and b.

a cycle. By the results of the previous section, $\log(\gamma_k)/k \to \log(\rho_1\rho_2)$. We consider two instances of this network both of which have $\rho_1 \times \rho_2 = 1/6$. Thus, the necessary condition for asymptotic optimality requires that $R = 6$. The two cases are as follows.

- Case I: $\lambda = 1$, $\mu_1 = 2$, $\mu_2 = 3$. Time reversal strongly suggests that the way A_k occurs is for queue 1 to first build up to a certain level greater than k, and then for queue 2 to build up while queue drains until A_k is hit near the point (k,k). Thus the intermediate sets A_j are inconsistent with this buildup behavior.
- Case II: $\lambda = 1$, $\mu_1 = 3$, $\mu_2 = 2$. This is the case in which the buildup to A_k occurs along the diagonal line from $(0,0)$ to (k,k) and thus the intermediate sets A_j are consistent with the large deviations behavior.

The simulations were run for 16 h on an IBM RS/6000 workstation, which resulted in approximately 12 billion transitions per run. From a single run at a given parameter setting, we simultaneously estimated γ_k for $k = 10, 20, 30, 40$. Splitting was done upon entry to the set A_j for $j = 1, \cdots, k$ and we used $R = 6$ splits per level, consistent with the necessary condition for optimality. By solving sets of linear equations, numerical values of γ_k can be computed thereby permitting comparison of the simulation estimates to numerically precise results. The results are reported in Table I, which displays γ_k, $\hat{\gamma}_k$ and the relative error (defined to be the relative width of a 99% confidence interval).

In Case I, notice that the relative error increases as k increases and that the estimate $\hat{\gamma}_k$ is orders of magnitude too low for large k, e.g., $\hat{\gamma}_{40}$ is five orders of magnitude off. The

TABLE I
SPLITTING RESULTS FOR THE TWO-QUEUE TANDEM JACKSON NETWORK

Case	k	γ_k	$\hat{\gamma}_k$	Relative Error
I	10	9.64×10^{-8}	9.82×10^{-8}	±7.4%
	20	1.60×10^{-15}	2.37×10^{-16}	±76%
	30	2.64×10^{-23}	6.21×10^{-26}	±104%
	40	4.36×10^{-31}	3.17×10^{-36}	±111%
II	10	9.64×10^{-8}	9.05×10^{-8}	±6.1%
	20	1.60×10^{-15}	1.33×10^{-15}	±16%
	30	2.64×10^{-23}	1.92×10^{-23}	±37%
	40	4.36×10^{-31}	2.50×10^{-31}	±46%

estimates are much better in Case II, but the facts that the relative error increases with k and that the associated 99% confidence interval for γ_{40} does not come close to containing γ_{40} are symptoms that the method is not asymptotically optimal. These experimental results are completely consistent with the theory developed in this paper. As a point of comparison, experimental results in [13] for the finite Markovian case (in which case splitting when done properly is asymptotically optimal) obtained relative errors of less than ±20% for $\gamma_k \approx 10^{-30}$ in less than 10 min of CPU time on the same RS/6000 workstation. The computer codes used in these two papers were also identical, save for the differences in generating the transitions of the embedded Markov chains.

VI. CONCLUDING REMARKS

We have developed necessary conditions for the effectiveness of multilevel splitting in estimating rare event probabilities. Our results address two settings—one in which the

number of levels remains fixed and the probabilities of moving from one to the next become small and another in which the number of levels increases while the probabilities remain fixed. In both settings we have emphasized the importance of choosing the levels in a way consistent with the most likely path to a rare set. Our results suggest that the apparent simplicity of multilevel splitting may be somewhat misleading. In the end, the effectiveness of the method depends critically on an understanding of the way a rare event occurs, much as an understanding of the large deviations of a process is central to importance sampling.

We briefly contrast the conclusions reached here with the more positive results in our earlier paper [13]. There, we identified classes of models where splitting provided asymptotically optimal estimates. In the settings we considered, the most important condition was that the number of splits per level be chosen correctly. Theorem 3 of this paper indicates that the sufficient conditions we gave are in fact necessary in far greater generality. But we have also seen that this necessary condition ceases to be sufficient without the additional model structure used in [13]. The models we treated there could be roughly described as having the property that an event can become rare essentially along just one dimension. As a consequence, the most likely path to the ultimate level cannot differ too greatly from the most likely path to an intermediate level, and it is precisely this that our necessary conditions seek to ensure.

ACKNOWLEDGMENT

This work was performed while T. Zajic was on a postdoctoral fellowship at Columbia University and IBM Research. The photo of P. Glasserman was taken by J. Pelaez.

REFERENCES

[1] S. Andradóttir, D. P. Heyman, and T. J. Ott, "On the choice of alternative measures in importance sampling with Markov chains," *Oper. Res.*, vol. 43, no. 3, pp. 509–519, 1995.

[2] V. Anantharam, "The optimal buffer allocation problem," *IEEE Trans. Inform. Theory*, vol. 35, pp. 721–725, 1989.

[3] J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation and Estimation*. New York: Wiley, 1990.

[4] C. S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 913–931, 1994.

[5] C. S. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin, "Effective bandwidth and fast simulation of ATM in tree networks," *Perf. Eval.*, vol. 20, pp. 45–65, 1994.

[6] C. S. Chang, P. Heidelberger, and P. Shahabuddin, "Fast simulation of packet loss rates in a shared buffer communications switch," *ACM Trans. Modeling and Computer Simulation*, vol. 5, no. 4, pp. 306–325, 1995.

[7] J. Collamore, "Hitting probabilities and large deviations," *Ann. Probability*, vol. 24, no. 4, pp. 2065–2078, 1996.

[8] M. Cottrell, J.-C. Fort, and G. Malgouyres, "Large deviations and rare events in the study of stochastic algorithms," *IEEE Trans. Automat. Contr.*, vol. 28, pp. 907–920, 1983.

[9] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. London, U.K.: Jones and Bartelett, 1993.

[10] M. Devetsikiotis and J. K. Townsend, "An algorithmic approach to the optimization of importance sampling parameters in digital communication system simulation," *IEEE Trans. Commun.*, vol. 41, no. 10, pp. 1464–1473, 1990.

[11] A. Dubi, "General statistical model for geometrical splitting in Monte Carlo—Part I," *Transport Theory and Statistical Phys.*, vol. 14, pp. 167–193, 1985.

[12] M. R. Frater, T. M. Lenon, and B. D. O. Anderson, "Optimally efficient estimation of the statistics of rare events in queueing networks," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 1395–1405, 1991.

[13] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic, "Multilevel splitting for estimating rare event probabilities," Yorktown Heights, NY, IBM T. J. Watson Res. Center Rep. RC 20478, 1996 and *Ops. Res.*, to be published.

[14] ——, "Splitting for rare event simulation: Analysis of simple cases," in *Proc. Winter Simulation Conf.* San Diego, CA: IEEE Computer Soc. Press, 1996, pp. 302–308.

[15] ——, "A look at multilevel splitting," in *Monte Carlo and Quasi-Monte Carlo Methods*, H. Niederreiter, P. Hellekalek, G. Larcher, and P. Zinterhof, Eds. New York: Springer-Verlag, 1997, pp. 98–108.

[16] P. Glasserman and S. G. Kou, "Analysis of an importance sampling estimator for tandem queues," *ACM Trans. Modeling and Computer Simulation*, vol. 5, no. 1, pp. 22–42, 1995.

[17] P. Glasserman and Y. Wang, "Counterexamples in importance sampling for large deviations probabilities," *Ann. Appl. Probability*, vol. 7, pp. 731–746, 1997.

[18] P. W. Glynn and W. Whitt, "The asymptotic efficiency of simulation estimators," *Oper. Res.*, vol. 40, pp. 505–520, 1992.

[19] A. Goyal, P. Shahabuddin, P. Heidelberger, V. F. Nicola, and P. W. Glynn, "A unified framework for simulating Markovian models of highly reliable systems," *IEEE Trans. Comput.*, vol. 41, pp. 36–51, 1992.

[20] J. Hammersley and D. Handscomb, *Monte Carlo Methods*. London, U.K.: Methuen, 1965.

[21] P. Heidelberger, "Fast simulation of rare events in queueing and reliability models," *ACM Trans. Modeling and Computer Simulation*, vol. 5, no. 1, pp. 43–85, 1995.

[22] H. Kahn and T. E. Harris, "Estimation of particle transmission by random sampling," *Nat. Bur. Standards Appl. Math. Series*, vol. 12, pp. 27–30, 1951.

[23] G. Kesidis and J. Walrand, "Quick simulation of ATM buffers with on-off multiclass Markov fluid sources," *ACM Trans. Modeling and Computer Simulation*, vol. 3, no. 3, pp. 269–276, 1993.

[24] S. Parekh and J. Walrand, "A quick simulation method for excessive backlogs in networks of queues," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 54–56, 1989.

[25] J. S. Sadowsky, "Large deviations and efficient simulation of excessive backlogs in a GI/G/m queue," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 1383–1394, 1991.

[26] ——, "On the optimality and stability of exponential twisting in Monte Carlo estimation," *IEEE Trans. Inform. Theory*, vol. 39, pp. 119–128, 1993.

[27] J. S. Sadowsky and J. A. Bucklew, "On large deviations theory and asymptotically efficient Monte Carlo estimation," *IEEE Trans. Inform. Theory*, vol. 36, pp. 579–588, 1990.

[28] F. Schreiber and C. Görg, "Rare event simulation: A modified RESTART-method using LRE-algorithm," in *Proc. Int. Telecommunications Conf.*, J. Labetoulle and J. W. Roberts, Eds. Amsterdam, The Netherlands: Elsevier Sci., 1994, pp. 787–796.

[29] P. Shahabuddin, "Importance sampling for the simulation of highly reliable Markovian systems," *Management Sci.*, vol. 40, no. 3, pp. 333–352, Mar. 1994.

[30] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis*. London, U.K.: Chapman & Hall, 1995.

[31] D. Siegmund, "Importance sampling in the Monte Carlo study of hypothesis tests," *Ann. Statist.*, vol. 4, pp. 673–684, 1976.

[32] A. V. Starkov, "Monte Carlo splitting importance sampling," *Monte Carlo Methods and Appl.*, vol. 1, pp. 241–250, 1995.

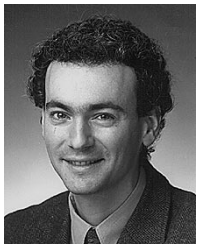
[33] G. de Veciana, C. Courcoubetis, and J. Walrand, "Decoupling bandwidths for networks: A decomposition approach to resource management," in *IEEE INFOCOM Proc.* New York: IEEE Computer Soc. Press, 1994, pp. 466–473.

[34] M. Villén-Altamirano, A. Martínez-Marrón, J. Gamo, and F. Fernández-Cuesta, "Enhancements of the accelerated simulation method RESTART by considering multiple thresholds," in *Proc. Int. Telecommunications Conf. 14*, J. Labetoulle and J. W. Roberts, Eds. Amsterdam, The Netherlands: Elsevier Sci., 1994, pp. 797–810.

[35] M. Villén-Altamirano and J. Villén-Altamirano, "RESTART: A method for accelerating rare event simulations," *Queueing, Performance and Control in ATM*, J. W. Cohen and C. D. Pack, Eds. Amsterdam, The Netherlands: Elsevier Sci., 1991, pp. 71–76.

[36] ——, "RESTART: A straightforward method for fast simulation of rare events," in *Proc. Winter Simulation Conf.* San Diego, CA: Society for Computer Simulation, 1994, pp. 282–289.

[37] T. Zajic, "On the typical path to hitting a rare set," in *Proc. 34th Annual Allerton Conf. Communication, Control and Computing*. Urbana-Champaign, IL: Univ. Illinois, 1996, pp. 751–759.



Paul Glasserman received the A.B. degree in mathematics from Princeton University, Princeton, NJ, in 1984 and the Ph.D. degree in applied mathematics from Harvard University, Cambridge, MA, in 1988.

He is a Professor in the Management Science division of the Columbia University Graduate School of Business and holds a secondary appointment in the Department of Industrial Engineering and Operations Research. Prior to joining the Columbia faculty, he was a Member of Technical Staff in the Operations Research department of AT&T Bell Laboratories.

Dr. Glasserman is a past recipient of the TIMS Outstanding Simulation Publication Award (1991), of an NSF National Young Investigator Award (1994), and of the INFORMS Erlang Prize in Applied Probability (1996).

Philip Heidelberger (M'82–SM'91–F'94) received the B.A. degree in mathematics from Oberlin College, Oberlin, OH, in 1974 and the Ph.D. degree in Operations Research from Stanford University, Stanford, CA, in 1978.

He has been a Research Staff Member at the IBM T. J. Watson Research Center in Yorktown Heights, NY since 1978. While on sabbatical in 1993–1994, he was a Visiting Scientist at Cambridge University and at ICASE, NASA Langley Research Center. His research interests include modeling and analysis of computer performance, probabilistic aspects of discrete-event simulations, and parallel simulation.

Dr. Heidelberger has won Best Paper Awards at the ACM SIGMETRICS and ACM PADS (Parallel and Distributed Simulation) Conferences and was twice awarded the INFORMS College on Simulation's Outstanding Publication Award. He has recently served as Editor-in-Chief of the ACM's *Transactions on Modeling and Computer Simulation.* He served as the Program Chairman of the 1989 Winter Simulation Conference and the Program Co-Chairman of the ACM SIGMETRICS/Performance'92 Conference. He is a Fellow of the ACM.

Perwez Shahabuddin (M'96) received the B.Tech degree in mechanical engineering (1984) from the Indian Institute of Technology, Delhi, followed by the M.S. degree in statistics (1987) and the Ph.D degree in operations research (1990) from Stanford University, Palo Alto, CA.

From 1990 to 1995 he was a Research Staff Member at the IBM T. J. Watson Research Center, Yorktown Heights, NY. Currently he is an Associate Professor at the Industrial Engineering and Operations Research Department at Columbia University, New York, NY. His research interests include stochastic modeling, discrete-event simulation and Monte Carlo techniques, and performance/reliability analysis of computer/communication systems.

Dr. Shahabuddin received the 1996 Outstanding Simulation Publication Award from the INFORMS College on Simulation, a 1996 CAREER Award from the National Science Foundation, and a 1997 Distinguished Faculty Teaching Award from the Columbia Engineering School Alumni Association. While at IBM, he received IBM External Honors Awards in 1992 and 1994, IBM Invention Achievement Award in 1995, and was one of the developers of the System Availability Estimator (SAVE) modeling tool. A paper based on his Ph.D. work was the winner of the first prize in the 1990 George E. Nicholson Student Paper Competition conducted by INFORMS. Currently he is serving as an Associate Editor for IEEE TRANSACTIONS ON RELIABILITY and is on the Editorial Board of *IIE Transactions-Operations Engineering*.

Tim Zajic received the Ph.D. degree in operations research from Stanford University, Stanford, CA, in 1994.

From 1994 to 1995 he was a visitor at the Centro de Investigación en Matemáticas (CIMAT) in Guanajuato, México, and subsequently, from 1995 to 1997, was an NSF Postdoctoral Fellow at the Department of Industrial Engineering and Operations Research at Columbia University and the IBM T. J. Watson Research Center. He is currently a Visiting Assistant Professor in the School of Mathematics at the University of Minnesota.