RC 20692 (01/14/97)
Computer Science

IBM Research Report

A Look At Multilevel Splitting

Paul Glasserman

Columbia Business School
New York, New York 10027

Philip Heidelberger

IBM Research Division
T.J. Watson Research Center
Yorktown Heights, New York 10598

Perwez Shahabuddin, Tim Zajic

IEOR Department
Columbia University
New York, New York 10027

IBM Research Division
Almaden · T.J. Watson · Tokyo · Zurich

A Look At Multilevel Splitting

Paul Glasserman
Columbia University

Philip Heidelberger
IBM Research

Perwez Shahabuddin
Columbia University

Tim Zajic
Columbia University and IBM Research

December 9, 1996

Abstract

This paper gives a non-technical overview of our work on the efficiency of a simple multilevel splitting technique for estimating rare event probabilities. For a particular class of stochastic models, this method is asymptotically optimal provided the splitting factor is properly chosen. However, for more general classes of processes, the method fails to be asymptotically optimal unless the splitting is done in a manner consistent with the dominant way in which the rare event happens (the large deviations behavior). In the absence of such consistency, the method also exhibits a kind of bias in which the estimate is too small with high probability. We briefly contrast our approach with one used in nuclear physics.

1 Introduction

Requirements for low cell loss rates in telecommunications networks and for low system failure rates in fault-tolerant computing systems have emphasized the importance of analyzing rare events in stochastic models. For example, in ATM (Asynchronous Transfer Mode) communications networks, quality of service requirements may be specified in terms of a cell loss rate in the range of 10^{-9}. Exact analysis is limited to models having special structure while numerical analysis is limited to models whose underlying state space is not too large. Asymptotic analysis, in the form of large deviations analysis, has become popular [1]. Large deviations deals with the asymptotic probability of a rare event and the way in which the rare event occurs. However large deviations typically only gives the rate at which a rare event probability approaches zero. Furthermore, the class of problems amenable to large deviations analysis is also limited. For example, in the realm of queueing networks, large deviations results typically only exist for single node networks (e.g., [3, 15, 24] and the references therein to name only a few), and for multiple node networks having special structure (e.g., [3]).

Simulation thus appears as an attractive alternative. However, efficient simulation techniques for estimating a rare event probability are essential since the run length requirement using standard simulation typically increases as the inverse of the probability being estimated. While many variance reduction techniques have been proposed, two seem particularly applicable to the rare event problem: importance sampling (e.g., [16, 13]) and splitting (e.g., [18]).

A survey on the application of importance sampling to problems arising in queueing and reliability models may be found in [17]. There is a relationship between large deviations and effective importance sampling. The zero variance importance sampling change of measure (page 58 of [16]) requires sampling from the conditional distribution of the rare event. Since the theory of large deviations describes the asymptotic distribution of the rare event, the change of measure suggested by large deviations appears to be a good candidate for use in importance sampling. Indeed this is the case in some simple cases [4, 5, 19, 20] such as a multiple server queue. However, in more complicated models heuristics suggested by large deviations may not only fail to be (asymptotically) efficient, but may in fact lead to variance inflation and sometimes infinite variance [11, 12]. Thus importance sampling needs to be applied with great caution in complex models, especially in complex queueing networks.

Various forms of splitting have existed in the literature since the early 1950's. It appears to be of widespread use in problems arising in nuclear physics, e.g., [6, 7, 2]. (We will contrast the use and analysis of these physics problems with our approach in Section 5.) However, it has only recently come into use for queueing and reliability problems [21, 22, 23], under the name of RESTART. On the surface, splitting is an easy method to implement: one needs only to decide when to split and what the splitting factor should be. This suggests that splitting should be a more broadly applicable procedure than importance sampling. Some approximate analysis of RESTART's efficiency has been performed, but that analysis does not relate to the large deviations behavior of the process being simulated.

In this paper, we summarize our work on splitting: detailed expositions can be found in [8, 9, 10]. To make this paper accessible to a larger audience, the description here will contain a minimum of notation, will leave a number of messy technical assumptions unstated, and will contain no proofs. Our particular form of splitting is described in Section 2. We then focus on when splitting is "asymptotically optimal." When a method is asymptotically optimal, its computing requirements grow slowly as the event becomes rarer. Indeed, as shown in Section 3, for a particular class of stochastic models, our form of splitting is asymptotically optimal provided the right splitting factor is selected. This optimal factor is identified for this class of problems. In Section 4, we relate the large deviations behavior of the process to the efficiency of splitting. We identify necessary conditions on the splitting factor and on the large deviations rate functions that must be satisfied in order for splitting to be asymptotically optimal. These conditions imply that the large deviations behavior of the process, i.e., the dominant way in which the rare event occurs, must be understood. Failure to be consistent with the process's large deviations behavior leads not only to a sub-optimal variance, but also to a form of bias in which the estimator may be orders of magnitude too low with high probability. Furthermore, even if the levels are consistent with the large deviations behavior, conditions on the rates of less likely paths to failure must also be satisfied to maintain asymptotic optimality. Such conditions, which have been shown to be violated for one of the simplest queueing network models, would be quite difficult to check for more complex models. Thus, as with importance sampling, the use of simple splitting techniques must be applied with caution.

2 Multilevel Splitting

We are interested in estimating γ_k which is defined to be the probability that a stochastic process, starting in some set A_0, reaches a "rare" set A_k before it returns to A_0. For example, in a queue, A_k may be the set where the total queue length exceeds k and efficient estimation of γ_k is key when estimating the steady-state buffer overflow rate, e.g., [17]. (Or more generally,

A_k may be the set where the total queue length exceeds some constant q_k.) As k increases, γ_k decreases and special simulation techniques become necessary. We assume throughout that the process being simulated is a countable state space Markov chain in either discrete or continuous time. We assume that there are intermediate sets (levels) $A_1, \ldots A_{k-1}$, constructed in such a way that the process must pass through A_{j-1} before reaching A_j. In this way, we can write $\gamma_k = p_1 \cdots p_k$ where p_j is the probability of reaching level j before hitting A_0 given that level $j-1$ was reached. We assume that γ_k approaches zero at rate $\rho^{k+o(k)}$ for some constant $0 < \rho < 1$, i.e., we assume

$$\lim_{k\to\infty} \log(\gamma_k)/k = \log(\rho). \tag{1}$$

To estimate γ_k, we use a simple multilevel splitting procedure as depicted in Figure 1. When the process first enters A_1, it splits into R_2 subpaths each of which is started from the entrance state of A_1. This procedure is repeated recursively whenever a subpath first enters the next highest level, i.e., upon first entrance to A_j, R_{j+1} new subpaths are started. (R_1 is the number of trials from the starting level A_0.) An unbiased estimate of γ_k is then

$$\hat{\gamma}_k = \frac{Z_k}{R_1 \cdots R_k} \tag{2}$$

where Z_k is the total number of subpaths that reach level k. Note that $\{Z_j, j \geq 0\}$ can be viewed as a branching process. We are interested in the asymptotic efficiency, as $k \to \infty$,

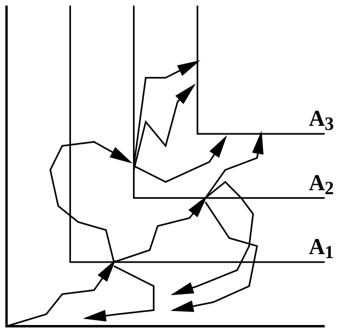

Figure 1: An example of multilevel splitting with two intermediate levels and a splitting factor of three. Two of the three subpaths from the first split reach the next level; four of their subpaths reach the final level.

of the method. Two factors affect the efficiency: w_k, the expected work per sample, and σ_k^2, the variance per sample. (For simplicity, we assume that the computational cost is one unit per subpath.) The appropriate measure of efficiency combines these into the work-variance product $w_k\sigma_k^2$ [14, 16]. Since, for any splitting factors R_j and intermediate levels A_j the mean is unchanged (i.e., $E[\hat{\gamma}_k] = \gamma_k$), we concentrate on $\eta_k = w_k \times E[\hat{\gamma}_k^2]$. With too little splitting, entrance to A_k is still a rare event, while with too much splitting, the work per sample grows but the second moment cannot be decreased below its minimum possible value γ_k^2. In fact, since $E[\hat{\gamma}_k^2] \geq \gamma_k^2$ and $w_k \geq 1$, there is a best possible rate at which η_k can approach zero and we will call the procedure asymptotically optimal if this best possible rate is achieved, i.e., if

$$\lim_{k\to\infty} \log(\eta_k)/k = 2\log(\rho). \tag{3}$$

3 The Good News

We now describe a setting in which splitting is asymptotically optimal under simple conditions [9]. We assume that the process being simulated is a two dimensional Markov chain with a state space denoted by (m, i). The first dimension denotes the process's level while the second dimension contains enough auxiliary information so as to render the process Markovian. For example, in an ATM queue fed by multiple Markov-modulated sources, m could represent the total queue length and i could represent an encoding that identifies the state of each source. We assume that there are only finitely many auxiliary states i. A_k then represents all states with first component $\geq k$.

We assume that there is a transition matrix $\mathbf{P}_m$ where $P_m(i, j)$ denotes the probability that the process first enters A_m in state j (before entering A_0) given that it first entered level $m-1$ in state i. Since the dynamics of the process may be very complicated, we do not assume that $\mathbf{P}_m$ is known in closed form, rather we only assume that samples can be simulated having this distribution. Because the probability of ever reaching the next highest level (before A_0) may be less than one, $\mathbf{P}_m$ may be a substochastic matrix. We assume that a limiting transition matrix $\mathbf{P}$ exists, i.e., $\lim_{m\to\infty} P_m(i, j) = P(i, j)$. We will refer to this class of models as "finite Markovian." Under appropriate irreducibility assumptions, (1) is satisfied with ρ being the spectral radius of $\mathbf{P}$. Furthermore, if a constant splitting factor is used at each level, i.e., if $R_j = R$ for all j, then the method is asymptotically optimal if and only if

$$R = 1/\rho. \tag{4}$$

With this choice of R, $E[Z_k]$ remains roughly constant. Since $1/\rho$ is typically not an integer, the constant number of splits from level j can be replaced by a random number of splits R_j, chosen independently from everything else, with mean $E[R_j]$. The estimator $\hat{\gamma}_k = Z_k/(E[R_1]\cdots E[R_k])$ is an unbiased estimate of γ_k and is asymptotically optimal if $E[R_j] = 1/\rho$ (and R_j is bounded).

Numerical results in [9] indicate that the method works extremely well when R is properly chosen. For example, in an ATM model with on-off sources, $\gamma_{40} = 2.33 \times 10^{-20}$ can be estimated to within $\pm 13\%$ error in only 500 seconds of CPU time on an IBM RS/6000 workstation. (Here, the relative error is the half-width of a 99% confidence interval divided by γ_k.) The performance of the method degrades as R deviates from $1/\rho$, although some deviation from optimality can be tolerated. In the above ATM example, the relative error increases to about $\pm 35\%$ when R is 20% greater than optimal. However, the method becomes more sensitive to deviations from $1/\rho$ as k increases and γ_k decreases. These results suggest that the simple splitting scheme described in Section 2, augmented by pilot studies to estimate a good splitting factor, should be highly effective for this broad class of problems.

4 The Bad News

Not all processes fit within the finite Markovian framework of the previous section (infinite in one dimension, finite in the other). For example, in a queueing network, many queues may be simultaneously large. In the finite Markovian case, $R = 1/\rho$ is a necessary and sufficient condition for asymptotic optimality. In this more general setting, assuming (1) holds, the condition $R = 1/\rho$, is a necessary, but not sufficient, condition for splitting to be asymptotically optimal [8].

Unlike the finite Markovian case, in which there is a natural set of intermediate levels and essentially only one way to reach A_k, there may now be many ways to reach A_k. Thus in designing the splitting procedure, not only must the splitting factors R_j be chosen, the intermediate levels A_j must also be chosen. As we will see, if these intermediate levels are not chosen properly, the method loses efficiency.

Consider the situation of Figure 2 in which some intermediate level B and the final level A are depicted. We assume that rare events in the process occur according to a large deviations

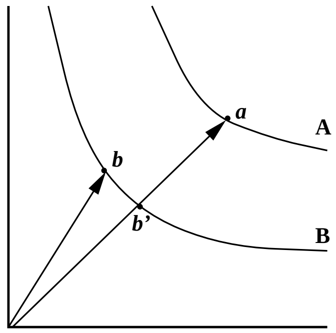

Figure 2: The most likely path to the intermediate set B hits B at b, but the most likely path to the final set A hits B at b'.

principle. Informally, given that the process enters B, it does so by following the path that ends at the point b with high probability. (By "following the path" we generally mean that an appropriately scaled version of the process stays close to such a deterministic path with probability approaching one.) Similarly, given that the process enters A, it does so by following the path that enters B at b' and ends at the point $a \in A$ with high probability.

Now consider the effect that such behavior has on the efficiency of splitting. Assume that the splitting factor is $R = 1/\rho$, in accordance with the necessary condition. By (1), (2) and the fact that $\hat{\gamma}_k$ is unbiased, $\log(E[Z_k])/k \to 0$. Almost all of the effort in simulating subpaths from level B will be spent on subpaths that originate at (or near) b since b is the most likely hitting point of B. However, since the most likely path to A does not pass through b, much of that effort is, in essence, wasted since few of those subpaths will reach the final level A. Thus $Z_k = 0$ with high probability. Only rarely, will a subpath enter B at the "correct" point b'. However, when such a subpath does enter b' it will eventually generate a large expected number of subpaths that enter A. This phenomenon precludes asymptotic optimality. Thus a necessary condition for asymptotic optimality is that $b = b'$, i.e., the large deviations path to any intermediate level must coincide with the large deviations path to the final level.

Furthermore, if the situation shown in Figure 2 exists a kind of bias occurs. More formally, there is a growth rate for the sample size $R_1 = R_1(k)$ of replications from level 0 and a set F_k such that $R_1(k) \to \infty$, $P\{F_k\} \to 1$ and

$$\lim_{k\to\infty} E[\hat{\gamma}_k 1_{F_k}]/\gamma_k = 0 \tag{5}$$

where 1_{F_k} is the indicator of the set F_k. The interpretation of (5) is as follows: for a large sample size $R_1(k)$, $\hat{\gamma}_k$ appears to underestimate γ_k by a large amount with high probability. Similar effects occur when the number of levels is fixed and the probabilities of moving to the next level approach zero.

However, even if the intermediate levels are consistent with large deviations as described above (i.e., $b = b'$), the method is not always guaranteed to be asymptotically optimal. Suppose the intermediate level B is a fraction α of the way to A, i.e., $\log(P\{B\})/k \to \alpha \log(\rho)$. Let x be any entrance point to B. Suppose the rate of hitting A by passing through x can be written as $\alpha \log(\delta_1(x)) + (1-\alpha)\log(\delta_2(x))$, i.e., the probability of hitting B near x is approximately $\delta_1(x)^{\alpha k}$ and the probability of going from x to A is approximately $\delta_2(x)^{(1-\alpha)k}$. For any such x,

$$\alpha \log(\delta_1(x)) + (1-\alpha)\log(\delta_2(x)) \leq \log(\rho) \tag{6}$$

since the path through x is not the only way to reach A. In order for splitting to be asymptotically optimal, an additional condition must be satisfied for every x and every intermediate level:

$$\alpha \log(\delta_1(x)) + 2(1-\alpha)\log(\delta_2(x)) \leq (2-\alpha)\log(\rho). \tag{7}$$

No matter how unlikely it is to enter B at x, (7) states that it cannot be too "easy" to reach A from x since (7) may be violated if $\delta_2(x) \gg \rho$. If (7) is violated, only infrequently will a subpath enter B at x, but when it does it will spawn an expected avalanche of subpaths that reach A.

The inequality (7) is a necessary, but not sufficient, condition for asymptotic optimality. Verification of (7) appears difficult in practice. However, it has been shown *not* to hold in a simple queueing network, even when the intermediate levels are consistent with large deviations. This network is a two queue tandem Jackson network (Poisson arrivals and exponential service time distributions) and the rare event of interest is when both queues are simultaneously large. Experimental results confirm that much slower convergence is obtained in this example as compared to finite Markovian examples. Also, in this example, if the intermediate levels are chosen in a way that is not consistent with large deviations, then the type of biasing predicted by (5) is observed experimentally: in one case $\hat{\gamma}_k$ is five orders of magnitude too low, even after hours of CPU time.

5 Contrast With Splitting in Physics

We now briefly contrast our approach to splitting with that described in [6, 7, 2] for studying problems in nuclear physics. There, the methods and analysis seem highly tailored to the particular problem setting, specifically estimating the probability that a particle passing through a material reaches a detector. The material is divided into cells and when a particle enters a new cell it may split (or be randomly killed). When a particle is split, its weight is adjusted: the weight contains information about both the splitting procedure and the physics of the problem. The splitting factor may be cell-dependent. In some splitting procedures, certain transitions may be forced to occur in which case the weight is adjusted by the probability of the transition. An approximation for the variance, σ^2, is derived. This expression is approximate for a number of reasons including the fact that an infinite series is truncated and an approximation is used that considers only the average behavior of particles passing through a cell (the "surface-point approximation"). An approximate expression for the work per sample, w, is also derived. The expressions for σ^2 and w are quite complicated and involve some unknown quantities that must be estimated via pilot studies. Such estimation is then embedded within a numerical optimization routine to minimize the estimated work-variance product $w\sigma^2$.

In our splitting approach, all subpaths reaching the final level carry the same weight, $1/R_1 \cdots R_k$, whereas in physics particles may carry different weights. This simplification makes asymptotic analysis possible. However, even with this simplification, in the finite Markovian setting our approach is as effective as *any* simulation algorithm can possibly be (in the sense of asymptotically minimizing the work times second moment). Furthermore, in this setting, the asymptotically optimal splitting factor can be simply identified using the mathematical structure of the problem, whereas in physics a numerical optimization procedure is applied. However, in more general settings the constant weight per subpath property of this simple approach may be a liability, since the method can easily fail to be asymptotically optimal. Whether or not large deviations analysis can be used to effectively assign state-dependent splitting factors (and thus weights) remains to be seen.

6 Acknowledgments

This work is supported by NSF grants DMI-94-57189, DMS-9508709 and DMI-96-25291.

References

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

[2] K. Burn. Optimizing cell importances using an extension of the DSA – theory, implementation, preliminary results. *Progress in Nuclear Energy*, 24:39–54, 1990.

[3] C.S. Chang. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Trans. Automatic Control*, 39:913–931, 1994.

[4] C.S. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin. Effective bandwidth and fast simulation of ATM intree networks. *Performance Evaluation*, 20:45–65, 1994.

[5] M. Cottrell, J.-C. Fort, and G. Malgouyres. Large deviations and rare events in the study of stochastic algorithms. *IEEE Trans. Automatic Control*, 28:907–920, 1983.

[6] A. Dubi. General statistical model for geometrical splitting in Monte Carlo – I. *Transport Theory and Statistical Physics*, 14:167–193, 1985.

[7] A. Dubi, A. Goldfeld, and K. Burn. Application of the direct statistical approach on a multisurface splitting problem in Monte Carlo calculations. *Nuclear Science and Engineering*, 93:204–213, 1986.

[8] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. A large deviations perspective on the efficiency of multilevel splitting. Technical report, IBM T.J. Watson Research Center, Yorktown Heights, NY, 1996.

[9] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Multilevel splitting for estimating rare event probabilities. Technical Report RC 20478, IBM T.J. Watson Research Center, Yorktown Heights, NY, 1996.

[10] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Splitting for rare event simulation: Analysis of simple cases. In *1996 Winter Simulation Conference Proceedings*, San Diego, CA, 1996.

[11] P. Glasserman and S.G. Kou. Analysis of an importance sampling estimator for tandem queues. *ACM Trans. Modeling and Computer Simulation*, 5:22–42, 1995.

[12] P. Glasserman and Y. Wang. Counterexamples in importance sampling for large deviations probabilities. Technical report, Columbia Business School, New York, NY, 1996.

[13] P.W. Glynn and D.L. Iglehart. Importance sampling for stochastic simulations. *Management Science*, 35:1367–1392, 1989.

[14] P.W. Glynn and W. Whitt. The asymptotic efficiency of simulation estimators. *Operations Research*, 40:505–520, 1992.

[15] P.W. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single server queue. *Journal of Applied Probability*, 31A:131–156, 1994.

[16] J.M. Hammersley and D.C. Handscomb. *Monte Carlo Methods.* Methuen and Co. Ltd., London, 1965.

[17] P. Heidelberger. Fast simulation of rare events in queueing and reliability models. *ACM Trans. Modeling and Computer Simulation*, 5:43 – 85, 1995.

[18] H. Kahn and T.E. Harris. Estimation of particle transmission by random sampling. *National Bureau of Standards Applied Mathematics Series*, 12:27–30, 1951.

[19] J.S. Sadowsky. Large deviations and efficient simulation of excessive backlogs in a GI/G/m queue. *IEEE Trans. Automatic Control*, 36:1383–1394, 1991.

[20] J.S. Sadowsky and J.A. Bucklew. On large deviations theory and asymptotically efficient Monte Carlo estimation. *IEEE Trans. Information Theory*, 36:579–588, 1990.

[21] M. Villén-Altamirano, A. Martínez-Marrón, J. Gamo, and F. Fernández-Cuesta. Enhancements of the accelerated simulation method RESTART by considering multiple thresholds. In J. Labetoulle and J.W. Roberts, editors, *International Telecommunications Conference 14*, pages 797–810, Amsterdam, 1994. Elsevier Science Publishers.

[22] M. Villén-Altamirano and J. Villén-Altamirano. RESTART: A method for accelerating rare event simulations. In J.W. Cohen and C.D. Pack, editors, *Queueing, Performance and Control in ATM*, pages 71–76. Elsevier Science Publishers, Amsterdam, 1991.

[23] M. Villén-Altamirano and J. Villén-Altamirano. RESTART: A straightforward method for fast simulation of rare events. In *1994 Winter Simulation Conference Proceedings*, pages 282–289, Orlando, FL, 1994.

[24] W. Whitt. Tail probability with statistical multiplexing and effective bandwidths in multiclass queues. *Telecommunication Systems*, 2:71–107, 1993.