Multilevel Splitting for Estimating Rare Event Probabilities

Paul Glasserman
Columbia University
New York, NY 10027

Philip Heidelberger
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598

Perwez Shahabuddin
Columbia University
New York, NY 10027

Tim Zajic
Columbia University
New York, NY 10027 *

1 Motivation

The estimation of rare event probabilities poses some of the of the most difficult computational challenges for Monte Carlo simulation and, at the same time, some of the greatest opportunities for efficiency improvement through the use of variance reduction techniques. Current interest in rare events stems primarily from developments in computer and communications technology: many industrial and scientific applications require highly reliable computer systems (with correspondingly small failure probabilities), and standards for emerging telecommunications systems call for extremely small buffer-overflow probabilities. The performance of these types of systems is frequently studied through simulation, but straightforward simulation can easily produce estimates that are off by orders of magnitude in estimating small probabilities. In these settings, variance reduction is essential.

Importance sampling, based on changing probability distributions to make rare events less rare, has been used to obtain dramatic improvements in efficiency in estimating small probabilities in queueing and reliability systems (see [4] and [7] for overviews). But the effectiveness of importance sampling depends critically on the ability to find the right change of measure; indeed, used improperly importance sampling is liable to produce worse results than straightforward simulation. Finding the right change of measure generally requires identifying at least the rough asymptotics of a rare event probability, often described by a large deviations result. This type of analysis can be formidable in complex models, so the domain of importance sampling, while substantial, does not include all problems of interest.

This work deals with an alternative method for rare event simulation that uses the technique of splitting sample paths. The main advantage of this technique is that it appears to require rather little model structure for its applicability. Splitting for rare event simulation was originally discussed by [6] in the context of estimating rare particle transmission probabilities in physics. Since then, there were only a few intermittent references to the use of this technique for rare event simulation ([2], [1], [5]). However, recently it was revisited in a significant way by [9], [8], and [10] for estimating probabilities of rare events in computer and communication systems. They also developed a software based modeling tool called ASTRO that implemented this method. Even though some approximate analysis of the efficiency of this method, that gives a few insights, has been done in the past, to date there does not exist a thorough formal analysis. The main purpose of this work is to describe a unifying class of models and implementation conditions under which this type of method is provably effective and even optimal (in an asymptotic sense) for rare event simulation. The analysis in this work takes extensively from the theory of branching processes (e.g., [3]).

2 The Technique and its Analysis

The method is best described through a simple example. Consider the simulation of a nonnegative process that returns to the origin infinitely often — think of the queue-length process in a stable queue. Consider the probability that, starting from the origin, the process reaches some level b before returning to the origin. As has been discussed in many past papers on rare event simulation, efficient estimation of this type of probability is central to efficient estimation of the steady-state probability that a queue length exceeds b (or the efficient estimation of the buffer overflow probability in a queueing system with finite buffer b).

*Also affiliated to IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

If b is large, this may well be a rare event; starting even a large number of sample paths at the origin may result in very few that reach b before returning, and thus generating little information about the probability of this event. To get around this problem we may partition the state space using intermediate thresholds as illustrated in Figure 1, where b corresponds to Level 3. Then, each time a sample path reaches a threshold higher than any it has reached before we split it into a number of subpaths, which subsequently evolve independently of each other. A path is terminated when it reaches level b or returns to the origin. Reaching an intermediate level is more

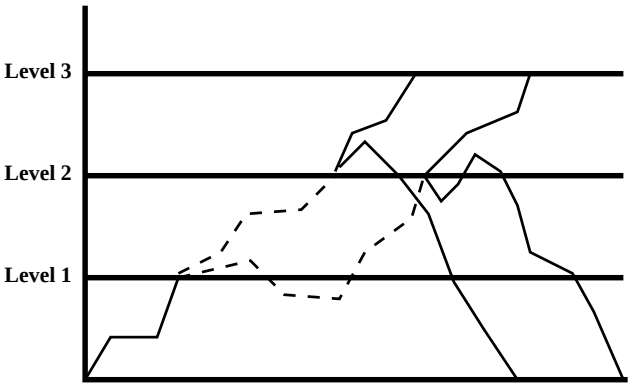

Figure 1: Splitting with three levels and two split subpaths.

likely than reaching b itself, and by splitting at each threshold we reinforce successful outcomes and end up allocating more effort to simulating more promising paths. Dividing the total number of paths that reach b before 0 by the total number of *potential paths* yields an unbiased estimate of the desired probability ([9] and [10] describe a slightly different implementation in which a path splits every time it crosses a threshold — even one it has reached before; [6] mentions both versions).

The central issues in implementing this method are choosing the thresholds and choosing the number of subpaths to generate when a path splits. In this work, we address only the second issue. Some of our conditions may be interpreted as roughly requiring that the thresholds be eventually nearly evenly spaced. More precisely, we will require that the dynamics of the process between thresholds approaches a limit at high thresholds. We plan to report results on the choice of threshold elsewhere, as that analysis involves rather different tools. Indeed, on more general state spaces than those we consider here the term "threshold" may be misleading; we require, in general, a nested sequence of subsets.

Our analysis in this work is based on modeling the movement from one threshold to the next rather than explicitly modeling the underlying process. Thus, our results may be viewed as an exact analysis of processes for which these models apply literally and an approximate analysis for more general cases. Briefly, we consider three settings allowing for increasing levels of generality:

- Upon crossing a threshold, the underlying process has a fixed *success* probability p of achieving the next threshold before terminating, independent of its past. Hence, the process that records the highest threshold reached so far is Markov. The requirement that the success probability be independent of the past holds if the underlying process is itself a Markov chain and there is a single entry state for each threshold. If, in addition, the underlying process is spatially homogeneous and the thresholds are evenly spaced, then the success probability is indeed constant.

- The process that records the highest threshold reached so far becomes homogeneous Markov when augmented with a supplementary variable taking on finitely many values. If, for example, the underlying process is Markov and the number of entry states per threshold is bounded, it suffices to record the highest threshold reached and the index of the state in which it was entered to get a Markov chain. In this setting, the movement from one threshold to the next is described by a matrix of transition probabilities.

- The movement from one threshold to the next is again described by transition probabilities, but we drop the requirement that a single transition matrix apply at all thresholds and replace it with the condition that the transition matrices converge to a limiting matrix.

The last setting is evidently the most general. For a specific example in which it applies, consider a queue in discrete time. Exactly one job is completed at each time increment so long as the system is not empty. Arrivals per time increment are i.i.d. and bounded. Take the underlying process to be the queue length and suppose the thresholds are at $\Delta, 2\Delta, 3\Delta, \ldots$ for some positive integer Δ larger than the greatest number of arrivals possible in a single time increment. Given that the queue length first achieved the threshold at $k\Delta$ by entering state $k\Delta + i$, for some $0 \leq i < \Delta$, the probability that it will achieve the next threshold (before returning to 0) by entering state $(k+1)\Delta + j$, for some $0 \leq j < \Delta$, is independent of the past. The movement from level k

to $k+1$ can thus be described by a $\Delta \times \Delta$ transition matrix with entries $P_k(i,j)$, and it is easy to see that these matrices converge as $k \to \infty$.

For each of the settings above we show that appropriately choosing the degree of splitting at each threshold is critical to the effectiveness of the method. The choice must balance two competing concerns: excessive splitting creates an explosive computational burden, and insufficient splitting eliminates the advantage over straightforward simulation. But with just the right amount of splitting, the method becomes *asymptotically optimal* (in a sense used frequently in rare event simulation) and is thus in some respects as effective for rare event simulation as any method can be. Our main results identify the ideal level of splitting for the three settings above: in the first setting, each path should be split into approximately $1/p$ subpaths; in the second setting the splitting parameter should be the reciprocal of the spectral radius of the transition matrix; and in the third setting it should be the reciprocal of the spectral radius of the limiting transition matrix. Often, this entails randomizing the number of subpaths. We obtain these results by modeling the paths that reach each threshold as the population at subsequent generations of a branching process. They may be loosely interpreted as stating that when a path splits, the number of subpaths should be chosen so that on average one subpath makes it to the next threshold. This keeps the expected number of paths alive at each threshold roughly constant.

In this work we analyze the three settings above. In addition we report numerical results on some simple computer system and communication network examples that support the theoretical analysis and explore the robustness of the method. Indeed, whereas the results of this work are essentially positive, it is important to emphasize that they are obtained under restrictions. Our purpose here is to show how well the method works under ideal conditions; elsewhere, we plan to address some of the limitations of the method, particularly in higher dimensional problems.

References

[1] Bayes, A.J. 1970. Statistical Techniques for Simulation Models. *Australian Computer J.* **2**, 180-184.

[2] Hammersley, J., and D. Handscomb. 1965. *Monte Carlo Methods.* Methuen & Co. Ltd., London.

[3] Harris, T. *The Theory of Branching Processes.* Dover, New York, 1989.

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

[5] Hopmans, A.C.M., and J.P.C. Kleijnen. 1979. Importance Sampling in Systems Simulation: A Practical Failure? *Math. and Computers in Simulation* **21**, 209–220.

[6] Kahn, H., and T.E. Harris. 1951. Estimation of Particle Transmission by Random Sampling. *National Bureau of Standards Applied Mathematics Series* **12**, 27–30.

[7] Shahabuddin, P. 1995. Rare Event Simulation in Stochastic Models. In *Proceedings of the 1995 Winter Simulation Conference*, 178–185, IEEE Computer Society Press.

[8] Villén-Altamirano, M., A. Martínez-Marrón, J. Gamo, and F. Fernández-Cuesta. 1994. Enhancements of the Accelerated Simulation Method RESTART by Considering Multiple Thresholds. *Proceedings of the 14th International Teletraffic Conference.* In *The fundamental role of teletraffic in the evolution of telecommunications networks*, J. Labetoulle and J.W. Roberts (eds.), 797–810. Elsevier Science Publishers, Amsterdam.

[9] Villén-Altamirano, M., and J. Villén-Altamirano. 1991. RESTART: A Method for Accelerating Rare Event Simulations. *Proceedings of the 13th International Teletraffic Congress.* In *Queueing, performance and control in ATM*, J.W. Cohen and C.D. Pack (eds.), 71–76. Elsevier Science Publishers, Amsterdam.

[10] Villén-Altamirano, M., and J. Villén-Altamirano. 1994. RESTART: A Straightforward Method for Fast Simulation of Rare Events. *Proceedings of the Winter Simulation Conference*, 282–289, Society for Computer Simulation, San Diego, California.