Asymptotic Robustness of Estimators in Rare-Event Simulation

PIERRE L'ECUYER, Université de Montreal, Canada
JOSE H. BLANCHET, Harvard University, USA
BRUNO TUFFIN, IRISA–INRIA, Rennes, France, and
PETER W. GLYNN, Stanford University, USA

The asymptotic robustness of estimators as a function of a rarity parameter, in the context of rare-event simulation, is often qualified by properties such as bounded relative error (BRE) and logarithmic efficiency (LE), also called asymptotic optimality. However, these properties do not suffice to ensure that moments of order higher than one are well estimated. For example, they do not guarantee that the variance of the empirical variance remains under control as a function of the rarity parameter. We study generalizations of the BRE and LE properties that take care of this limitation. They are named bounded relative moment of order k (BRM-k) and logarithmic efficiency of order k (LE-k), where $k \geq 1$ is an arbitrary real number. We also introduce and examine a stronger notion called vanishing relative centered moment of order k, and exhibit examples where it holds. These properties are of interest for various estimators, including the empirical mean and the empirical variance. We develop (sufficient) Lyapunov-type conditions for these properties in a setting where state-dependent importance sampling (IS) is used to estimate first-passage time probabilities. We show how these conditions can guide us in the design of good IS schemes, that enjoy convenient asymptotic robustness properties, in the context of random walks with light-tailed and heavy-tailed increments. As another illustration, we study the hierarchy between these robustness properties (and a few others) for a model of highly-reliable Markovian system (HRMS) where the goal is to estimate the failure probability of the system. In this setting, for a popular class of IS schemes, we show that BRM-k and LE-k are equivalent and that these properties become strictly stronger when k increases. We also obtain a necessary and sufficient condition for BRM-k in terms of quantities that can be readily verified from the parameters of the model.

Categories and Subject Descriptors: G.3 [**Probability and Statistics**]: Probabilistic algorithms (including Monte Carlo); I.6.1 [**Simulation and Modeling**]: Simulation Theory

General Terms: Algorithms, Performance

Additional Key Words and Phrases: Rare-event simulation, robustness, bounded relative error, logarithmic efficiency, importance sampling, zero-variance approximation

Authors' addresses: Pierre L'Ecuyer, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, H3C 3J7, Canada; email: `lecuyer@iro.umontreal.ca`; Jose H. Blanchet, Department of Statistics, Harvard University, Cambridge, MA 02138, USA, email: `blanchet@fas.harvard.edu`; Bruno Tuffin, IRISA-INRIA, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France, email: `Bruno.Tuffin@irisa.fr`; Peter W. Glynn, Department of Management Science and Engineering, Stanford University, Stanford, CA 94305-4026, USA, email: `glynn@stanford.edu`.

1. INTRODUCTION

Rare-event simulation refers to the situation where a set of events that occur very rarely in a simulation model are important and must be taken into account because their occurrence have high consequences. It is a key tool for decision making in several areas such as reliability, telecommunications, finance, insurance, and computational chemistry and physics, among others [Bolhuis et al. 2002; Bucklew 2004; Heidelberger 1995; Juneja and Shahabuddin 2006; Kalos and Whitlock 1986]. The important rare events may correspond, for example, to huge financial losses, or environmental disasters, or loss of lives, or other types of accidents. Before we decide on how much money we want to spend (or what additional measures we want to take) to avoid these rare events, we need to have an idea of their probability of occurrence and of the effect of additional spending on this probability.

In typical rare-event settings, the Monte Carlo method is not viable unless special "acceleration" techniques are used to make the important rare events occur frequently enough for moderate sample sizes. The two main families of techniques for doing that are splitting [Ermakov and Melas 1995; Glasserman et al. 1998; L'Ecuyer et al. 2007; Villén-Altamirano and Villén-Altamirano 2006] and importance sampling (IS) [Bucklew 2004; Glynn and Iglehart 1989; Heidelberger 1995; Juneja and Shahabuddin 2006].

Suppose we want to estimate a positive quantity $\gamma = \gamma(\varepsilon)$ that depends on a *rarity* parameter $\varepsilon > 0$. We assume that $\lim_{\varepsilon \to 0^+} \gamma(\varepsilon) = 0$. We have a family of estimators $Y = Y(\varepsilon)$ taking their values in $[0, \infty)$, such that $\mathbb{E}[Y(\varepsilon)] = \gamma(\varepsilon) > 0$ for each $\varepsilon > 0$. In applications, $\gamma(\varepsilon)$ can be a performance measure defined as a mathematical expectation, and some model parameters are defined as functions of ε in a convenient way. Note that this parameterization by ε is introduced only for the asymptotic analysis of estimators. Different parameterizations may correspond to different asymptotic regimes. For example, in a queuing system for which we are interested in the probability that the queue length exceeds a given (large) threshold B, we may take $\varepsilon = 1/B$ to study what happens when B gets larger and larger. If we are interested in the behavior of the queue for a large number s of servers, we may take $\varepsilon = 1/s$. In other settings, the service time and inter-arrival time distributions might depend on ε. In Markovian reliability models, the failure rates and repair rates might be functions of ε. For example, when studying a highly-reliable system where the failure rates are very small, the failure rates are often taken as polynomial functions of ε for the purpose of asymptotic analysis [Nakayama 1996; Shahabuddin 1994].

The convergence speed of $\gamma(\varepsilon)$ toward 0 may depend on how the model is parameterized, but the robustness properties introduced in this paper do not depend on this speed; they depend only on the magnitude of certain moments of $Y(\varepsilon)$ relative to the corresponding powers of $\gamma(\varepsilon)$.

A special case of this setting arises when $Y(\varepsilon)$ is an indicator function: $Y(\varepsilon) = 1$ with probability $\gamma(\varepsilon)$ and $Y(\varepsilon) = 0$ with probability $1 - \gamma(\varepsilon)$. In this case, $\mathrm{Var}[Y(\varepsilon)] = \gamma(\varepsilon)(1-\gamma(\varepsilon)) \approx \gamma(\varepsilon)$, so the squared relative error (or relative variance) $\mathrm{Var}[Y(\varepsilon)]/\gamma^2(\varepsilon) \approx 1/\gamma(\varepsilon)$ grows without bound when $\varepsilon \to 0$. If we estimate $\gamma(\varepsilon)$ by the average of $n = n(\varepsilon)$ independent copies of $Y(\varepsilon)$, we have an estimator with relative variance $1/(n(\varepsilon)\gamma(\varepsilon))$. This estimator does not have bounded relative error

(BRE) unless the sample size $n(\varepsilon)$ grows at least at the same rate as $1/\gamma(\varepsilon)$ when $\varepsilon \to 0$ [Heidelberger 1995], which means that the computing budget would have to increase without bound. Viewed from another angle, if we fix the computing budget to a constant, so $n(\varepsilon)$ is not allowed to grow indefinitely when $\varepsilon \to 0$, then the relative error is unbounded.

In this type of situation, splitting and IS are often used to design better estimators, which may have the BRE property with a fixed computing budget. There are many cases (e.g., in queueing and finance) where the best available estimators do not have the BRE property, but enjoy the slightly weaker property of logarithmic efficiency (LE), also called asymptotic optimality. This often happens when the estimators are constructed by exploiting the theory of large deviations [Asmussen 2002; Glasserman 2004; Heidelberger 1995; Juneja and Shahabuddin 2006; Siegmund 1976]. LE has the intuitive interpretation that when $\gamma^2(\varepsilon) \to 0$ exponentially fast in $1/\epsilon$, $\mathrm{Var}[Y(\varepsilon)] \to 0$ at the same exponential rate.

To see why the BRE or LE properties are often not sufficient, suppose we want to compute a confidence interval on $\gamma(\varepsilon)$ based again on independent replicates of $Y(\varepsilon)$. To do this via the classical central limit theorem (CLT), we need reliable estimators for both the mean $\gamma(\varepsilon)$ and the variance $\sigma^2(\varepsilon) = \mathbb{E}[(Y(\varepsilon) - \gamma(\varepsilon))^2]$. We want these estimators to remain robust in the sense that their relative error remains bounded (or grows only very slowly) when $\epsilon \to 0$. Under the assumption that one uses a confidence interval with a half-width proportional to the exact (theoretical) variance, the relative half-width remains bounded if the estimator has BRE [Heidelberger 1995]. But to realistically implement such a confidence interval procedure, one needs to estimate the variance from the simulated i.i.d. runs of the model. To obtain such a confidence interval, in which the relative half-width is estimated properly, one typically needs an estimator of $\sigma^2(\varepsilon)$ that is accurate to order $\gamma^2(\varepsilon) \times o(1)$ as $n \to \infty$, uniformly in ε. Obtaining a variance estimator with such a level of relative accuracy (relative to $\gamma^2(\varepsilon)$) requires control over the $(2+\delta)$th moment of $Y(\varepsilon)$ for some $\delta > 0$. In rare-event settings, reliable (relative) mean and variance estimators are typically difficult to obtain. In fact, the *relative variance* is often more difficult to estimate than the mean (relative to the mean).

A similar problem arises in empirically comparing the efficiencies of two different estimators for the quantity $\gamma(\varepsilon)$, as $\varepsilon \to 0$. In particular, the efficiency is typically assessed by comparing the variances of the associated estimators. Since the exact (theoretical) variances are not available analytically, they must be computed from the sample variance, as obtained from the simulation runs used to estimate $\gamma(\varepsilon)$. Even if all the estimators to be compared enjoy the BRE property, a potentially huge number of simulation runs may be required to compute the ratio of efficiencies between the available estimators, unless the fourth moment of the estimator scales in proportion to $\gamma^4(\varepsilon)$.

This motivates our introduction, in this paper, of asymptotic characterizations that generalize BRE and LE, namely *bounded relative moment of order k* (BRM-k) and *logarithmic efficiency of order k* (LE-k), where $k \in [1, \infty)$. The relative moment of order k is the expectation of $[Y(\varepsilon)/\gamma(\varepsilon)]^k$. An estimator has the BRM-k property if its relative moment of order k remains bounded when $\varepsilon \to 0$. The LE-k property roughly means that when $\gamma^k(\varepsilon) \to 0$ at an exponential rate, the

kth moment converges to zero at the same exponential rate. BRE-2 and LE-2 are equivalent to BRE and LE, respectively. We also introduce and discuss a much stronger property than BRM-k, named *vanishing relative centered moment of order k (VRCM-k), which means that the relative centered moment of order k converges to 0 when* $\varepsilon \to 0$. *As it turns out, this property implies that the sampling scheme converges to a zero-variance sampling scheme when* $\varepsilon \to 0$. *We give examples where this property holds.*

These concepts apply to *any estimator* that depends on some rarity parameter ε; it does not have to involve splitting or IS. This includes for instance the empirical variance and higher empirical moments taken as estimators of the exact variance and of higher moments of the estimator of interest. For example, saying that the *empirical variance* has the BRM-2 property means that the variance of the empirical variance, divided by the squared variance, is bounded when $\varepsilon \to 0$. This is *bounded relative error of the empirical variance* (as a variance estimator). Saying that the empirical mean has the BRM-4 property, on the other hand, means that its fourth moment divided by the fourth power of the mean is bounded. These two properties are not equivalent in general.

Lesser-known asymptotic robustness properties than BRE and LE have also been studied in the literature. For instance, Sadowsky [1993] examines a generalization of LE for central empirical moments of high-order, in a specific large-deviations context where the goal is to estimate the probability that the average of $n = \lfloor 1/\varepsilon \rfloor$ i.i.d. random variables exceeds a given constant. Boots and Shahabuddin [2000] define a weaker criterion than LE, motivated by the observation that the large variance sometimes comes from a set of events with "small" probability relative to the probability of the rare event itself, uniformly in ε. If the restriction of the estimator to the large set (defined as the complement of this set of small probability) is LE, they say that the estimator has *large set asymptotic optimality*. Other properties include *bounded normal approximation* (BNA), and *asymptotic good estimation of the mean* (AGEM) and *of the variance* (AGEV) (also called *probability* and *variance well-estimation*) [Tuffin 1999; 2004]. BNA, as defined in Tuffin [1999], implies that if we approximate the distribution of the average of n i.i.d. copies of Y by the normal distribution (e.g., to compute a confidence interval), the approximation is accurate to order $O(n^{-1/2})$ uniformly in ε when $\varepsilon \to 0$. AGEM and AGEV have been defined in the context of estimating a probability in a highly reliable Markovian system (HRMS), and basically mean that the sample paths that contribute the most to the estimator and its second moment, respectively, are not rare under the sampling scheme that is examined.

It is important to underline that all notions mentioned so far completely disregard the computational work (CPU time) required to obtain the estimator. In general, this computational cost can be random, and its mean or higher moments, which often depends on ε, can be unbounded when $\varepsilon \to 0$. This motivates the need for *work-normalized* versions of the BRM-k, LE-k, and VRCM-k properties. For $k = 2$, the standard practice for taking the work into account when comparing estimators is to multiply the variance by the expected computational cost [Hammersley and Handscomb 1964; Glynn and Whitt 1992], based on the idea that doubling the computing budget typically permits one (roughly) to halve the

variance. This has motivated the introduction of concepts such as *bounded work-normalized relative error* (also called *bounded relative efficiency*) in Cancela et al. [2005] and *work-normalized logarithmic efficiency* (or asymptotic optimality) in Boots and Shahabuddin [2000] and Glasserman et al. [1999], simply by multiplying the variance by the expected computing time in the definitions of BRE and LE. One could think of straightforward generalizations to any $k \geq 1$: just multiply the centered moments by the expected computing time. But this normalization is not necessarily appropriate, for a number of reasons. For example, if we have an estimator defined as an average over n independent replications, doubling the number of replications does not divide the kth centered moment by 2 in general, for $k \neq 2$. Even for $k = 2$, a concept that considers only the expected computing time would not guarantee that we can compute a reliable confidence interval for $\gamma(\varepsilon)$ uniformly over ε, for a given large computing budget that does not depend on ε. If the (random) computing time has unbounded moments of order larger than 1 when $\varepsilon \to 0$, then for any fixed computing budget c, the probability of completing at least one replication within the budget limit may go to zero when $\varepsilon \to 0$, for example. Thus, just multiplying by the expected computing time does not necessarily provide the desired notion of boundedness; it could even be misleading to some extent. For these reasons, we end our discussion of work-normalization here and leave this important topic for another paper.

It is important to recognize that estimators with a higher level of robustness do not necessarily require a larger computational effort. A well-designed IS scheme often reduces the simulation time by pushing the system faster toward the rare event, while decreasing higher moments at the same time, so we may win on both fronts: smaller moments and a smaller computing time. For instance, as we shall discuss in Section 4.1, importance sampling estimators designed to have either the LE-2 or the BRE property often satisfy the corresponding improved measures of robustness such as LE-k and BRM-k for $k > 2$ as well. In Section 5, the more robust estimators are not really more expensive to compute either.

After defining and discussing the robustness properties, we examine some specific rare-event settings in which we study the relationships between these properties and provide easily verifiable conditions for these properties to hold.

Our basic setting is a discrete-time Markov chain (DTMC) model for which we want to estimate the probability $\gamma(x, \varepsilon)$ of reaching B before A in finite time, where A and B are two disjoint subsets of the state space, and the chain starts in state $x \notin A \cup B$. Either B, or the transition kernel of the DTMC, or both, may depend on ε. We focus on a general class of state-dependent IS schemes that attempt to approximate the zero-variance IS scheme for this model. The zero-variance IS scheme simply multiplies the transition probability (or density) from a state x to another state y by the product $\gamma(y, \varepsilon)/\gamma(x, \varepsilon)$. In practice, the function $\gamma(\cdot)$ is unknown (otherwise there would be no need to simulate in the first place), but if we replace its use in the construction of the zero-variance IS scheme by an approximation of good quality as $\varepsilon \to 0$, a significant accuracy improvement can often be achieved. The chain is simulated under the modified probability laws obtained from the approximation, and the original estimator is multiplied (as usual) by an appropriate weight called the likelihood ratio, to counter-balance the bias

caused by the change of measure. This type of state-dependent IS has been the focus of substantial research in both heavy-tailed and light-tailed settings during recent years (see, for instance, Dupuis and Wang [2004], Dupuis and Wang [2005] and Blanchet and Glynn [2007]). The approximation of $\gamma(\cdot)$ is usually obtained via large deviations theory or heavy-tailed approximation. One has to be careful, though: even with a good approximation in most of the state space, the likelihood ratio may sometimes exhibit a poor behavior due to the contributions corresponding to areas where the asymptotic description is not good enough.

In our DTMC setting, we establish general sufficient conditions for the BRM-k, LE-k, and VRCM-k properties. These conditions can be verified in terms of a simple Lyapunov inequality that involves the approximation of $\gamma(\cdot)$ together with some appropriate Lyapunov function. We apply these conditions for the design of IS estimators that exhibit BRM-k or LE-k, for random walks with both light-tailed and heavy-tailed increments. We also make the connection with other results found in the literature, e.g., by Sadowsky [1993] and Dupuis and Wang [2004], and we extend the results of the latter authors.

We then examine the robustness properties for an HRMS model studied by several authors [Cancela et al. 2002; Goyal et al. 1992; Heidelberger 1995; Lewis and Böhm 1984; Nakayama 1996; Shahabuddin 1994; Tuffin 1999; 2004], and used for reliability analysis of computer and telecommunication systems. In this model, a smaller value of the rarity parameter ε implies a smaller failure rate for the system's components, and we want to estimate the probability that the system reaches a "failed" state before it returns to a state where all the components are operational. This probability converges to 0 when $\varepsilon \to 0$. The model fits the DTMC setting mentioned earlier. For this HRMS model, specific conditions on the model parameters and on the IS probabilities have been obtained for the BRE property [Nakayama 1996], for BNA [Tuffin 1999; 2004], and for AGEM and AGEV [Tuffin 2004]. It is also shown by Tuffin [2004] that BNA implies AGEV, which implies BRE, which implies AGEM, which implies BRE, and that for each implication the converse is not true. In this paper we extend this hierarchy to incorporate BRM-k and LE-k, showing that for these models, these properties are all equivalent for any given k. We also obtain a necessary and sufficient condition on the model parameters for these properties to hold, for a given class of IS measures that covers all interesting IS schemes developed in the literature for these HRMS models. These conditions turn out to be of strictly increasing strength as a function of k. That is, if they hold for $k+1$ then they hold for k, but the converse is false for all k. We do this not only for the mean estimator, but for the estimators of all higher moments as well.

The remainder of the paper is organized as follows. In Section 2, we give formal definitions of the asymptotic characterizations discussed so far, along with simple examples. The main results of that section are Propositions 2.19 and 2.21; they prove the equivalence between two definitions of VRCM-k and the fact that VRCM-k implies convergence toward a zero-variance sampling scheme.

In Section 3, we define the Markov chain setting in which we want to estimate the probability of reaching B before A. We discuss the zero-variance approximation, we prove an upper bound on the kth moment under an IS scheme based on this

approximation and assuming a Lyapunov condition (Proposition 3.1), and we use this bound to derive sufficient conditions for BRM-k and for LE-k in this setting (Theorem 3.2). In Section 4, we use these conditions to study state-dependent IS estimators in random walks with light and heavy-tailed increments. Sections 4.1 and 4.2 introduce the model and recall what is known for state-independent IS when estimating the probability that the average of $n = \lfloor 1/\varepsilon \rfloor$ i.i.d. light-tailed random variables exceeds a given threshold. One can obtain LE-k but not BRM-k. In Section 4.3, we define a state-dependent IS scheme and prove in Proposition 4.5 that it has the BRM-k property. In Section 4.4, Theorem 4.6 extends a result of Dupuis and Wang [2004] and provides a sufficient condition for LE-k in the context of multidimensional random walks. In Section 4.5, we develop an IS scheme for the case of heavy-tailed distributions and show in Theorem 4.8 that it has the BRM-k property. In Section 5, we describe the HRMS model and we study the asymptotic robustness properties for a class of IS estimators applied to this model. For a large class of IS schemes, Theorem 5.2 gives necessary and sufficient conditions for BRM-k for the empirical moment of any order $g \geq 1$, and Proposition 5.6 shows the equivalence between LE-k and BRM-k. Proposition 5.5 also shows that this class of IS schemes cannot provide VRCM-k estimators. For a slightly different class of IS estimators, we prove in Proposition 5.7 that BRM-2 for the empirical variance implies BNA, then we provide a counterexample showing that the converse is not true.

We use the following notation. For a function $f : (0, \infty) \to \mathbb{R}$, we say that $f(\varepsilon) = o(\varepsilon^d)$ if $f(\varepsilon)/\varepsilon^d \to 0$ as $\varepsilon \to 0$; $f(\varepsilon) = O(\varepsilon^d)$ if $|f(\varepsilon)| \leq c_1\varepsilon^d$ for some constant $c_1 > 0$ for all ε sufficiently small; $f(\varepsilon) = \underline{O}(\varepsilon^d)$ if $|f(\varepsilon)| \geq c_2\varepsilon^d$ for some constant $c_2 > 0$ for all ε sufficiently small; and $f(\varepsilon) = \Theta(\varepsilon^d)$ if $f(\varepsilon) = \underline{O}(\varepsilon^d)$ and $f(\varepsilon) = O(\varepsilon^d)$. We use the shorthand notation $Y(\varepsilon)$ to refer to the family of estimators $\{Y(\varepsilon), \varepsilon > 0\}$. We also write "$\to 0$" to mean "$\to 0^+$."

2. ASYMPTOTIC ROBUSTNESS PROPERTIES

This section collects all the definitions, together with simple examples and counterexamples. The main novel results are in Section 2.6.

2.1 Bounded relative moments

DEFINITION 2.1. For $k \in [1, \infty)$, the *relative moment of order k* of the estimator $Y(\varepsilon)$ is defined as

$$m_k(\varepsilon) = \mathbb{E}[Y^k(\varepsilon)]/\gamma^k(\varepsilon). \tag{1}$$

The *variance* is

$$\sigma^2(\varepsilon) = \mathrm{Var}[Y(\varepsilon)] = \mathbb{E}[(Y(\varepsilon) - \gamma(\varepsilon))^2],$$

the *relative variance* is $\sigma^2(\varepsilon)/\gamma^2(\varepsilon)$, and the *relative error* is $\sigma(\varepsilon)/\gamma(\varepsilon)$.

DEFINITION 2.2. The estimator $Y(\varepsilon)$ has a *bounded relative moment of order k* (BRM-k) if

$$\limsup_{\varepsilon \to 0} m_k(\varepsilon) < \infty. \tag{2}$$

It has *bounded relative variance*, or equivalently *bounded relative error* (BRE) [Hei-

delberger 1995], if

$$\limsup_{\varepsilon\to 0} \sigma(\varepsilon)/\gamma(\varepsilon) < \infty. \tag{3}$$

EXAMPLE 2.3. Suppose $Y(\varepsilon)$ has a Pareto distribution with density $f(y) = a(\varepsilon)/y^{a(\varepsilon)+1}$ for $y > 1$, and $a(\varepsilon) = k_0 - \varepsilon$ for some integer $k_0 \geq 2$. In this case, for $k < k_0 - \varepsilon$, $\mathbb{E}[Y^k(\varepsilon)] = a(\varepsilon)/(a(\varepsilon) - k)$. Then, if $k < k_0$ and ε is small enough,

$$\frac{\mathbb{E}[Y^k(\varepsilon)]}{\gamma^k(\varepsilon)} = \frac{(k_0 - 1 - \varepsilon)^k}{(k_0 - k - \varepsilon)(k_0 - \varepsilon)^{k-1}},$$

so $Y(\varepsilon)$ is BRM-k.

EXAMPLE 2.4. It is shown in Bourin and Bondon [1998] that if $Y_j = X^j/\mu^j$ where $\mu_j = \mathbb{E}[X^j]$, j is a positive integer, and X is a non-negative random variable, then the variance of Y_j is non-decreasing in j. This implies that if $Y_j(\varepsilon) = X^j(\varepsilon)/\mu^j(\varepsilon)$ has the BRM-2 property, then $Y_{j'}(\varepsilon)$ also has it for all $j' < j$.

When computing a confidence interval on $\gamma(\varepsilon)$ based on the average of n i.i.d. replications of $Y(\varepsilon)$ and the (classical) central-limit theorem, for a fixed confidence level, the width of the confidence interval is (approximately) proportional to the standard deviation $\sigma(\varepsilon)$ divided by $\sqrt{n}$. Usually, the confidence interval has the form $(Y(\varepsilon) \pm z_{1-\alpha/2}\hat{\sigma}(\varepsilon)n^{-1/2})$, where $1-\alpha$ is the confidence level, $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution, and $\hat{\sigma}(\varepsilon)$ is the square root of the empirical variance of $Y(\varepsilon)$. The BRE property means that this width decreases at least as fast as $\gamma(\varepsilon)$ when $\varepsilon \to 0$.

It would perhaps seem natural to replace "$\limsup_{\varepsilon\to 0}$" in this definition by "$\sup_{0<\varepsilon\leq 1}$" for example. The definition would then be a bit stronger, so VRCM-k would no longer imply BRM-k, for example. We think that the difference is just a technicality that is not important in typical applications.

PROPOSITION 2.5. *BRE is equivalent to BRM-2.*

PROOF. This follows from the fact that $m_2(\varepsilon) = \mathbb{E}[Y^2(\varepsilon)]/\gamma^2(\varepsilon) = 1+\sigma^2(\varepsilon)/\gamma^2(\varepsilon)$. □

More generally, an equivalent definition of BRM-k is obtained if we replace $m_k(\varepsilon)$ in (2) by the *relative centered moment* $c_k(\varepsilon)$, defined by

$$c_k(\varepsilon) = \frac{\mathbb{E}[|Y(\varepsilon) - \gamma(\varepsilon)|^k]}{\gamma^k(\varepsilon)} = \mathbb{E}\left[\left|\frac{Y(\varepsilon)}{\gamma(\varepsilon)} - 1\right|^k\right]. \tag{4}$$

The equivalence follows from the following proposition:

PROPOSITION 2.6. *For any $k \geq 1$,*

$$\limsup_{\varepsilon\to 0} c_k(\varepsilon) < \infty \quad \textit{if and only if} \quad \limsup_{\varepsilon\to 0} m_k(\varepsilon) < \infty. \tag{5}$$

PROOF. We have

$$|Y(\varepsilon) - \gamma(\varepsilon)|^k \leq [\max(Y(\varepsilon), \gamma(\varepsilon))]^k \leq Y^k(\varepsilon) + \gamma^k(\varepsilon)$$

and

$$Y^k(\varepsilon) \leq [2\max(|Y(\varepsilon) - \gamma(\varepsilon)|, \gamma(\varepsilon))]^k \leq 2^k[|Y(\varepsilon) - \gamma(\varepsilon)|^k + \gamma^k(\varepsilon)],$$

from which

$$|Y(\varepsilon) - \gamma(\varepsilon)|^k \geq 2^{-k} Y^k(\varepsilon) - \gamma^k(\varepsilon).$$

Combining these inequalities, we obtain that

$$2^{-k} m_k(\varepsilon) - 1 \leq c_k(\varepsilon) \leq m_k(\varepsilon) + 1$$

and the result follows. □

PROPOSITION 2.7. *For any fixed ε and $k \geq 1$, $m_k(\varepsilon)$ is nondecreasing in k.*

PROOF. Since $Y(\varepsilon) \geq 0$, this follows from Jensen's inequality: if $1 \leq k' < k$, then

$$m_{k'}(\varepsilon) = \frac{\mathbb{E}[Y^{k'}(\varepsilon)]}{\gamma^{k'}(\varepsilon)} \leq \frac{(\mathbb{E}[(Y^k(\varepsilon))])^{k'/k}}{\gamma^{k'}(\varepsilon)} = \frac{\mathbb{E}[Y^k(\varepsilon)]}{\gamma^k(\varepsilon)} \frac{\gamma^{k-k'}(\varepsilon)}{(\mathbb{E}[(Y^k(\varepsilon))])^{(k-k')/k}} \leq m_k(\varepsilon).$$

□

COROLLARY 2.8. *BRM-k implies BRM-k' for $1 \leq k' < k$.*

Note that Proposition 2.7 would not hold if BRM-k was defined using the *centered* moment $\mathbb{E}[(Y(\varepsilon) - \gamma(\varepsilon))^k]$ instead of the non-centered moment $\mathbb{E}[Y^k(\varepsilon)]$ or the absolute centered moment $\mathbb{E}[|Y(\varepsilon) - \gamma(\varepsilon)|^k]$. This is illustrated by the following example.

EXAMPLE 2.9. Suppose $Y(\varepsilon)$ has the normal distribution with mean and variance $\gamma(\varepsilon) = \sigma^2(\varepsilon) = \varepsilon$. Then, $\mathbb{E}[(Y(\varepsilon) - \gamma(\varepsilon))^2]/\gamma^2(\varepsilon) = \sigma^2(\varepsilon)/\gamma^2(\varepsilon) = 1/\varepsilon$, whereas $\mathbb{E}[(Y(\varepsilon) - \gamma(\varepsilon))^3]/\gamma^3(\varepsilon) = 0$.

The following property is sometimes useful.

PROPOSITION 2.10. *For any positive real numbers k, ℓ, m, and any non-negative random variable $X(\varepsilon)$, if $Y(\varepsilon) = X^\ell(\varepsilon)$ is BRM-mk, then $Y'(\varepsilon) = X^{m\ell}(\varepsilon)$ is BRM-k.*

PROOF. From Jensen's inequality, $(\mathbb{E}[X^\ell(\varepsilon)])^{mk} \leq (\mathbb{E}[X^{m\ell}(\varepsilon)])^k$. Then,

$$\frac{\mathbb{E}[(X^{m\ell}(\varepsilon))^k]}{(\mathbb{E}[X^{m\ell}(\varepsilon)])^k} \leq \frac{\mathbb{E}[X^{mk\ell}(\varepsilon)]}{(\mathbb{E}[X^\ell(\varepsilon)])^{mk}} = \frac{\mathbb{E}[(X^\ell(\varepsilon))^{mk}]}{(\mathbb{E}[X^\ell(\varepsilon)])^{mk}}. \tag{6}$$

□

2.2 Logarithmic efficiency

There are several rare-event applications where practical BRE estimators are not readily available (e.g., in queueing and finance), but where estimators with the (weaker) LE property have been constructed by exploiting the theory of large deviations [Asmussen 2002; Glasserman 2004; Heidelberger 1995; Juneja and Shahabuddin 2006; Siegmund 1976]. Often, these estimators turn out to have the following LE-k property for all k.

DEFINITION 2.11. The estimator $Y(\varepsilon)$ is LE-k if

$$\lim_{\varepsilon \to 0} \frac{\ln \mathbb{E}[Y^k(\varepsilon)]}{k \ln \gamma(\varepsilon)} = 1. \tag{7}$$

LE-k means that when $\gamma^k(\varepsilon)$ converges to zero exponentially fast, $\mathbb{E}[Y^k(\varepsilon)]$ also converges exponentially fast and at the same exponential rate. This is the best possible rate; it cannot converge at a faster rate because from Jensen's inequality, we always have $\mathbb{E}[Y^k(\varepsilon)]-\gamma^k(\varepsilon) \geq 0$. LE-2 is the usual definition of LE, also known under the names of asymptotic efficiency and asymptotic optimality. In general, LE-k is weaker than BRM-k. But there are situations where the two are equivalent; this will happen in our HRMS setup in Section 5. The following examples illustrate the two possibilities. They correspond to the two types of parameterizations most often used in rare-event asymptotic analysis: The probability of the rare event decreases exponentially with ε in one case and polynomially in the other case. The exponential case typically occurs in situations where $\gamma(\varepsilon)$ satisfies a large deviations principle. The polynomial case is standard in HRMS models, for example, where the $\gamma(\varepsilon) \to 0$ because the transitions leading to the rare event have probabilities that decrease polynomially when $\varepsilon \to 0$, while their number remains fixed. We will return to this type of situation in Example 2.23 and in Section 5.

EXAMPLE 2.12. Suppose that $\gamma(\varepsilon) = q(\varepsilon)\exp[-\eta/\varepsilon]$ for some polynomial function q and some constant $\eta > 0$, and that our estimator has $\sigma^2(\varepsilon) = \exp[-2\eta/\varepsilon]$. Then, the LE property is easily verified, whereas BRE does not hold because $m_2(\varepsilon) = 1/q(\varepsilon) + 1 \to \infty$ when $\varepsilon \to 0$. We will see concrete examples of this situation in Section 4.

EXAMPLE 2.13. Suppose that $\gamma^k(\varepsilon) = q_1(\varepsilon) = \varepsilon^{t_1} + o(\varepsilon^{t_1})$ and $\mathbb{E}[Y^k(\varepsilon)] = q_2(\varepsilon) = \varepsilon^{t_2} + o(\varepsilon^{t_2})$. That is, both converge to 0 as a polynomial in ε. Clearly, $t_2 \leq t_1$, because $\mathbb{E}[Y^k(\varepsilon)]-\gamma^k(\varepsilon) \geq 0$. We have BRM-$k$ if and only if (iff) $q_2(\varepsilon)/q_1(\varepsilon)$ remains bounded when $\varepsilon \to 0$, iff $t_2 = t_1$. On the other hand, $-\ln q_1(\varepsilon) = -\ln(\varepsilon^{t_1}(1 + o(1))) = -t_1 \ln(\varepsilon) - \ln(1 + o(1))$ and similarly for $q_2(\varepsilon)$ and t_2. Then,

$$\lim_{\varepsilon\to 0} \frac{\ln \mathbb{E}[Y^k(\varepsilon)]}{k \ln \gamma(\varepsilon)} = \lim_{\varepsilon\to 0} \frac{t_2 \ln \varepsilon}{t_1 \ln \varepsilon} = \frac{t_2}{t_1}.$$

Thus, LE-k holds iff $t_2 = t_1$, which means that BRM-k and LE-k are equivalent in this case.

2.3 Bounded Normal Approximation

We mentioned earlier the computation of a confidence interval on $\gamma(\varepsilon)$ based on the central-limit theorem. This type of confidence interval is reliable if the sample average has approximately the normal distribution, so it is relevant to examine the quality of this normal approximation when $\varepsilon \to 0$. An error bound for this approximation is provided by the following generalization of the Berry-Esseen inequality [Bentkus and Götze 1996], first proved by Katz [1963]

THEOREM 2.14. *(Berry-Esseen) Let $Y_1, \ldots, Y_n$ be i.i.d. random variables with mean 0, variance σ^2, and third absolute moment $\beta_3 = \mathbb{E}[|Y_1|^3]$. Let $\bar{Y}_n$ and S_n^2 be the empirical mean and variance of $Y_1, \ldots, Y_n$, and let F_n denote the distribution function of the standardized sum (or Student statistic)*

$$S_n^* = \sqrt{n}\bar{Y}_n/S_n.$$

Then, there is an absolute constant $a < \infty$ such that for all $x \in \mathbb{R}$ and all $n \geq 2$,

$$|F_n(x) - \Phi(x)| \leq \frac{a\beta_3}{\sigma^3\sqrt{n}},$$

where Φ is the standard normal distribution function.

Note that the classical result usually has σ in place of S_n in the definition of S_n^* [Feller 1971]. Theorem 2.14 motivated the introduction by Tuffin [1999] of the BNA property, which requires that the Berry-Esseen bound remains $O(n^{-1/2})$ when $\varepsilon \to 0$.

DEFINITION 2.15. The estimator $Y(\varepsilon)$ has the *bounded normal approximation* (BNA) property if

$$\limsup_{\varepsilon \to 0} \frac{\mathbb{E}\left[|Y(\varepsilon) - \gamma(\varepsilon)|^3\right]}{\sigma^3(\varepsilon)} < \infty. \tag{8}$$

This BNA property *implies* that $\sqrt{n}|F_n(x) - \Phi(x)|$ remains bounded as a function of ε, i.e., that the approximation of F_n by the normal distribution remains accurate up to order $O(n^{-1/2})$, uniformly in ε. The reverse is not necessarily true, however. It may seem more natural to *define* the BNA property as meaning that $\sqrt{n}|F_n(x) - \Phi(x)|$ remains bounded, but Definition 2.15 has already been adopted in other papers mainly because it is often easier to obtain necessary and sufficient conditions for BNA with this definition.

If a confidence interval of level $1 - \alpha$ is obtained using the normal distribution while the true distribution is F_n, the error of coverage of the computed confidence interval does not exceed $2\sup_{x\in\mathbb{R}} |F_n(x) - \Phi(x)|$. If that confidence interval is computed from an i.i.d. sample $Y_1(\varepsilon), \ldots, Y_n(\varepsilon)$ of $Y(\varepsilon)$, BNA implies that the coverage error remains in $O(n^{-1/2})$ when $\varepsilon \to 0$, with a hidden constant that does not depend on ε.

BNA is not equivalent to BRM-3, because we divide by $\sigma^3(\varepsilon)$ in the definition of BNA and by $\gamma^3(\varepsilon)$ for BRM-3. One can have BNA and not BRM-3 (or BRM-3 and not BNA) if $\gamma(\varepsilon)$ converges to zero faster than $\sigma(\varepsilon)$ (or the opposite). If $\sigma(\varepsilon) = \Theta(\gamma(\varepsilon))$, then the two properties are equivalent.

Note that there are more general versions of the Berry-Esseen inequality that require only a bounded moment of order $2+\delta$ for any $\delta \in (0, 1]$ instead of the third moment β_3; see Petrov [1995, Theorem 5.7]. However, the bound on $|F_n(x) - \Phi(x)|$ in that case converges only as $O(n^{-\delta/2})$ instead of $O(n^{-1/2})$.

2.4 Asymptotic good estimation of the mean and of the variance

AGEM and AGEV are two additional robustness properties introduced by Tuffin [2004], under the name of "well estimated mean and variance," in the context of the application of IS to an HRMS model. Here we provide more general definitions of these properties. We assume that $Y(\varepsilon)$ is a *discrete* random variable, which takes value y with probability $p(\varepsilon, y) = \mathbb{P}[Y(\varepsilon) = y]$, for $y \in \mathbb{R}$. We also assume that its mean and variance are polynomial functions of ε: $\gamma(\varepsilon) = \Theta(\varepsilon^{t_1})$ and $\sigma^2(\varepsilon) = \Theta(\varepsilon^{t_2})$ for some constants $t_1 \geq 0$ and $t_2 \geq 0$. AGEM and AGEV state that the sample paths that contribute to the highest-order terms in these polynomial functions are not rare.

DEFINITION 2.16. (AGEM and AGEV) The estimator $Y(\varepsilon)$ has the AGEM property if $yp(\varepsilon, y) = \Theta(\varepsilon^{t_1})$ implies that $p(\varepsilon, y) = \Theta(1)$ (or equivalently, that $y = \Theta(\varepsilon^{t_1})$). It has the AGEV property if $[y - \gamma(\varepsilon)]^2 p(\varepsilon, y) = \Theta(\varepsilon^{t_2})$ implies that $p(\varepsilon, y) = \Theta(1)$ (or equivalently, that $[y - \gamma(\varepsilon)]^2 = \Theta(\varepsilon^{t_2})$).

These properties mean that for the realizations y of Y that provide the leading contributions to the estimator, the contributions decrease only because of decreasing values of y, and not because of decreasing probabilities. In a setting where IS is applied and Y is the product of an indicator function by a likelihood ratio (this will be the case in Sections 5.2 and 5.3), this means that the value of the likelihood ratio when $yp(\varepsilon, y)$ contributes to the leading term must converge at the same rate at this leading term when $\varepsilon \to 0$.

2.5 Robustness of the empirical variance

An important special case that we now examine is the stability of the empirical variance as an estimator of the true variance $\sigma^2(\varepsilon)$. Let $X_1(\varepsilon), \ldots, X_n(\varepsilon)$ be an i.i.d. sample of $X(\varepsilon)$, where $n \geq 2$. The empirical mean and empirical variance are $\bar{X}_n(\varepsilon) = (X_1(\varepsilon) + \cdots + X_n(\varepsilon))/n$ and

$$S_n^2(\varepsilon) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i(\varepsilon) - \bar{X}_n(\varepsilon))^2.$$

If we take $Y(\varepsilon) = S_n^2(\varepsilon)$ in our framework of the previous subsections, we obtain definitions of the robustness properties for $S_n^2(\varepsilon)$ as an estimator of $\sigma^2(\varepsilon)$. Let $\gamma(\varepsilon) = \mathbb{E}[X(\varepsilon)]$ (not $\mathbb{E}[Y(\varepsilon)]$ for now).

PROPOSITION 2.17. *If $\sigma^2(\varepsilon) = \Theta(\gamma^2(\varepsilon))$, then BRM-2k for $X(\varepsilon)$ implies BRM-k for $S_n^2(\varepsilon)$, for any $k \geq 1$.*

PROOF. Under the given assumption,

$$\frac{\mathbb{E}[S_n^{2k}(\varepsilon)]}{\sigma^{2k}(\varepsilon)} \leq \frac{\mathbb{E}[X^{2k}(\varepsilon)]}{\sigma^{2k}(\varepsilon)} = \Theta\left(\frac{E[X^{2k}(\varepsilon)]}{\gamma^{2k}(\varepsilon)}\right).$$

□

The BRM-4 property for a given estimator $X(\varepsilon)$ and the BRE property for its corresponding empirical variance $S_n^2(\varepsilon)$ are both linked to its fourth moment, so we might think that they are equivalent. In fact, we know (e.g., [Wilks 1962, page 200] or [Kendall and Stuart 1977, Exercise 10.13]) that

$$\mathrm{Var}[S_n^2(\varepsilon)] = \frac{1}{n}\left(\mathbb{E}[(Y(\varepsilon) - \mathbb{E}[Y(\varepsilon)])^4] - \frac{n-3}{n-1}\sigma^4(\varepsilon)\right). \tag{9}$$

Therefore,

$$\frac{\mathrm{Var}[S_n^2(\varepsilon)]}{\sigma^4(\varepsilon)} = \Theta\left(\frac{\mathbb{E}[(X(\varepsilon) - \gamma(\varepsilon))^4]}{\sigma^4(\varepsilon)}\right)$$

which differs in general from

$$\Theta\left(\frac{\mathbb{E}[X^4(\varepsilon)]}{\gamma^4(\varepsilon)}\right).$$

Thus, BRM-4 for $X(\varepsilon)$ and BRE for $S_n^2(\varepsilon)$; they are not equivalent in general. For example, $\sigma^2(\varepsilon)$ may converge to zero either at a faster rate or at a slower rate than $\gamma^2(\varepsilon)$. If $\sigma^2(\varepsilon) = \Theta(\gamma^2(\varepsilon))$ and $\mathbb{E}[(Y(\varepsilon) - \gamma(\varepsilon))^4] = \Theta(\mathbb{E}[Y^4(\varepsilon)])$, then they are equivalent. A similar observation applies to the equivalence between LE-4 for $X(\varepsilon)$ and LE for $S_n^2(\varepsilon)$ are not equivalent in general.

2.6 Vanishing relative centered moments

There are situations where not only the relative moment of order k is bounded, but its centered version also converges to zero when $\varepsilon \to 0$. We will give examples of that. It turns out that when this happens for any moment of order larger than 1, we are sampling asymptotically (as $\varepsilon \to 0$) from a zero-variance distribution.

DEFINITION 2.18. The estimator $Y(\varepsilon)$ has *vanishing relative centered moment of order k* (VRCM-k) if

$$\limsup_{\varepsilon \to 0} c_k(\varepsilon) = 0. \tag{10}$$

It has *vanishing relative variance*, or equivalently *vanishing relative error* (VRE), if

$$\limsup_{\varepsilon \to 0} \frac{\sigma(\varepsilon)}{\gamma(\varepsilon)} = 0. \tag{11}$$

Obviously, VRCM-k implies VRCM-k' for $1 \le k' \le k$, and similarly for the work-normalized versions. The following gives an equivalent definition of VRCM-k:

PROPOSITION 2.19. *For any $k \ge 1$,*

$$\limsup_{\varepsilon \to 0} m_k(\varepsilon) = 1 \quad \textit{if and only if} \quad \limsup_{\varepsilon \to 0} c_k(\varepsilon) = 0. \tag{12}$$

To prove this result we will use the following lemma:

LEMMA 2.20. *For any $k > 1$ and $\delta \in (0, k-1)$, there is a constant $A(\delta) > 0$ such that for all $x \ge 0$,*

$$\delta |x - 1| \le x^k - kx + (k-1) + A(\delta). \tag{13}$$

Moreover, $A(\delta)$ can be chosen so that $A(\delta) = \Theta(\delta^2)$ as $\delta \to 0$.

PROOF. Fix $\delta > 0$ and suppose first that $x \ge 1$. Consider the function

$$f_+(x) = x^k - (k+\delta)x + (k-1) + \delta.$$

Note that $f_+'(x_+(\delta)) = 0$ implies $x_+(\delta) = ((k+\delta)/k)^{1/(k-1)} > 0$. Since f_+ is strictly convex, we conclude that $f_+(x_+(\delta)) < 0$ is the global minimum of f_+. Therefore, we conclude that for all $x \ge 1$

$$\delta(x-1) \le x^k - kx + (k-1) - f_+(x_+(\delta)).$$

Now, observe that

$$\begin{aligned} f_+(x_+(\delta)) &= \left(1 + \frac{\delta}{k}\right)^{k/(k-1)} - (k+\delta)\left(1 + \frac{\delta}{k}\right)^{1/(k-1)} + (k-1) + \delta \\ &= 1 + \frac{\delta}{(k-1)} + \Theta(\delta^2) - (k+\delta)\left(1 + \frac{\delta}{k(k-1)}\right) + (k-1) + \delta \end{aligned}$$

$$= \Theta\left(\delta^2\right)$$

as $\delta \to 0$. A completely analogous strategy can be applied to the function

$$f_-(x) = x^k - (k-\delta)\,x + (k-1) - \delta$$

for $x \in [0,1)$, in which case we have that the minimizer is $x_-(\delta) = ((k-\delta)/k)^{1/(k-1)}$ with $\Theta\left(\delta^2\right) = f_-(x_-(\delta)) < 0$. We can then conclude that (13) holds with $A(\delta) = -[f_-(x_-(\delta)) + f_+(x_+(\delta))] = \Theta\left(\delta^2\right)$. □

Proof of Proposition 2.19. First we show that $\limsup_{\varepsilon\to 0} m_k(\varepsilon) = 1$ must imply that $\limsup_{\varepsilon\to 0} c_k(\varepsilon) = 0$. Applying Lemma 2.20 with $x = Y(\varepsilon)/\gamma(\varepsilon)$, taking expectations and $\varepsilon \to 0$, we find that

$$\limsup_{\varepsilon\to 0} \mathbb{E}\left[|Y(\varepsilon)/\gamma(\varepsilon) - 1|\right] \le A(\delta)/\delta.$$

Then we let $\delta \to 0$ and conclude that $Y(\varepsilon)/\gamma(\varepsilon) \to 1$ in the $\mathcal{L}_1$ norm and, in particular, in probability. Since, the random variables $Y^k(\varepsilon)/\gamma^k(\varepsilon)$ are non-negative and their expectation converges to unity as $\varepsilon \to 0$, then we must have uniform integrability and therefore convergence of $Y(\varepsilon)/\gamma(\varepsilon)$ in the $\mathcal{L}_k$ norm as $\varepsilon \to 0$ [Durrett 1996, page 260]. For the converse implication, the assumption that $\limsup_{\varepsilon\to 0} c_k(\varepsilon) = 0$ for $k > 1$ implies both convergence in probability to unity and uniform integrability of the random variables $Y^k(\varepsilon)/\gamma^k(\varepsilon)$. This implies in turn that $\limsup_{\varepsilon\to 0} m_k(\varepsilon) = 1$. □

Suppose we want to estimate

$$\gamma(\varepsilon) = \mathbb{E}_{P_\varepsilon}[Y(\varepsilon)] = \int_\Omega Y(\varepsilon,\omega) dP_\epsilon(\omega)$$

for some probability measure P_ϵ that depends on ε and some non-negative random variable $Y(\varepsilon)$, where Ω is the sample space. We may think of P_ε as the probability law that we are using to simulate our model. It could be the law of a Markov chain, for example, and it may include some variance reduction strategies such as importance sampling, splitting, and so on. In this context, we have a zero-variance change of measure with the new measure Q^*_ε defined by

$$\frac{dQ^*_\varepsilon}{dP_\varepsilon}(\omega) = \frac{Y(\varepsilon,\omega)}{\gamma(\varepsilon)}.$$

Recall that the total variation distance between two measures P and Q is defined by $|P-Q|_\infty = \sup_A |P(A) - Q(A)|$, where the sup is over all measurable sets.

Proposition 2.21. *If $Y(\varepsilon)$ is VRCM-$(1+\delta)$ for some $\delta > 0$, then $|P_\varepsilon - Q^*_\varepsilon|_\infty = o(1)$.*

Proof. Assuming that A runs over all measurable subsets of Ω, we have

$$\begin{aligned}
\sup_A |P_\varepsilon(A) - Q^*_\varepsilon(A)| &\le \sup_A |\mathbb{E}_{P_\varepsilon}\left[(dQ^*_\varepsilon/dP_\varepsilon)\,\mathbb{I}(A)\right] - \mathbb{E}_{P_\varepsilon}[\mathbb{I}(A)]| \\
&\le \mathbb{E}_{P_\varepsilon}\left|dQ^*_\varepsilon/dP_\varepsilon - 1\right| \\
&\le \mathbb{E}^{1/(1+\delta)}_{P_\varepsilon}\left[\left|dQ^*_\varepsilon/dP_\varepsilon - 1\right|^{(1+\delta)}\right]
\end{aligned}$$

$$\leq \mathbb{E}_{P_\varepsilon}^{1/(1+\delta)}\left[|Y(\varepsilon)/\gamma(\varepsilon)-1|^{(1+\delta)}\right]$$
$$= o(1). \qquad \square$$

In Proposition 2.21, we may have that only P_ε is a function of ε and not Y, or only Y and not $P_\varepsilon \equiv P$, or both are functions of ε. This proposition indicates that a VRCM-k estimator (with $k > 1$) based on importance sampling induces a distribution that is close (in total variation) to the zero-variance sampler, and even converges to it when $\varepsilon \to 0$. This might suggest that the design of such an estimator in situations of practical interest is hopeless. However, simulation schemes have recently been shown to achieve VRCM-k for $k > 1$ in some situations where a zero-variance IS scheme is used in which the exact function γ is replaced by an approximation v that converges to γ uniformly when $\varepsilon \to 0$ [L'Ecuyer and Tuffin 2008a; 2008b]. This happens for instance in the general Markov chain model examined in Example 2.23 below, which can be encountered in various situations, including reliability settings such as the HRMS models discussed in Section 5 and in L'Ecuyer and Tuffin [2008b]. The class of sampling schemes examined in Section 5 do not satisfy conditions (14) and (15), but it is possible to design a sampling scheme that does satisfy these conditions, along the lines of Example 2.23, and the corresponding estimator will then be VRCM-k. Other examples where a VRCM-k property holds in queueing and insurance problems can be found in Blanchet and Glynn [2007] and Juneja [2007].

Note that in a Markov chain setting, the probability of reaching a given set of states B (where the rare event occurs) can be small either because reaching B requires a large number of "upstream" transitions (and that number increases when $\varepsilon \to 0$), or because all sample paths leading to B have transitions whose probabilities are very small (and converge to 0 when $\varepsilon \to 0$) while the number of transitions may remain bounded. The following two examples illustrate how VRCM-k can be achieved (or not) in this second case. We start with a small concrete illustration; then we show in Example 2.23 how the results can be extended to a general class of Markov chain models.

Example 2.22. This small example gives a concrete illustration where a simple change of the transition probabilities can provide VRCM-k. Consider a system with two types of components and two components of each type. It evolves as a DTMC $\{X_j,\, j \geq 0\}$ whose state $X_j = (X_j^{(1)}, X_j^{(2)})$ at step j gives the number of failed components of each type. The system is down (in failure state) when the two components of any given type are down, i.e., when its state belongs to the set $B = \{(0,2),(1,2),(2,2),(2,1),(2,0)\}$. We want to estimate the probability $\gamma(\varepsilon)$ that a system starting in state $x_0 = (0,0)$ reaches B before it returns to state x_0. For this, we simulate this chain using IS by replacing the transition probabilities $p(x,y,\varepsilon) = \mathbb{P}[X_j = y \mid X_{j-1} = x]$ by new probabilities $q(x,y,\varepsilon)$. The probabilities $p(x,y,\varepsilon)$ and $q(x,y,\varepsilon)$ are given in Table I, in which the five states of B are aggregated in a single state called B.

Let Π_B be the set of sample paths $\pi = (x_0, x_1, \ldots, x_\tau)$ going from x_0 to B, where $\tau = \min\{j : x_j \in B\}$. Each path π has probability $p(\pi,\varepsilon) = \prod_{j=1}^{\tau} p(x_{j-1}, x_j, \varepsilon)$. The most likely path leading to B is $\pi_1 = ((0,0),(1,0),B)$ and its probability is $(1-\varepsilon^{12})\varepsilon^6 = \varepsilon^6 + O(\varepsilon^{18})$. It is not difficult to see that we also have $\gamma(\varepsilon) = \varepsilon^6 + o(\varepsilon^6)$

Table I. Transition probabilities for Example 2.22; the entry in row x and column y gives the original transition probability $p(x, y, \varepsilon)$ from state x to state y (top) and the modified probability $q(x, y, \varepsilon)$ (bottom).

	(0,0)	(0,1)	(1,0)	(1,1)	B
(0,0)		ε^{12}	$1-\varepsilon^{12}$		
		ε^{12}	$1-\varepsilon^{12}$		
(0,1)	$1-\varepsilon^2-\varepsilon^4$			ε^4	ε^2
	0			ε^2	$1-\varepsilon^2$
(1,0)	$1-\varepsilon^6-\varepsilon^8$			ε^8	ε^6
	0			ε^2	$1-\varepsilon^2$
(1,1)		$1/2-\varepsilon^4$	$1/2-\varepsilon^4$		$2\varepsilon^4$
		1/4	1/4		1/2

Table II. Values of $b(\pi)$, $c(\pi)$, and $\delta(k, \pi)$ (for $k = 2, 3, 4$), for each acyclic path in Π_B.

Path π	$b(\pi)$	$c(\pi)$	$\delta(2,\pi)$	$\delta(3,\pi)$	$\delta(4,\pi)$
$((0,0),(0,1),B)$	14	12	4	0	-4
$((0,0),(0,1),(1,1),B)$	20	14	14	14	14
$((0,0),(0,1),(1,1),(1,0),B)$	20	14	14	14	14
$((0,0),(1,0),B)$	6	0	0	0	0
$((0,0),(0,1),(1,1),B)$	12	2	10	14	18
$((0,0),(0,1),(1,1),(1,0),B)$	12	2	10	14	18

(the next example gives a proof in a more general setting). When we reach B via some path $\pi \in \Pi_B$, the estimator $Y(\varepsilon)$ takes the value $p(\pi, \varepsilon)/q(\pi, \varepsilon)$, which is the corresponding likelihood ratio, and this happens with probability $q(\pi, \varepsilon)$. Note that $p(\pi, \varepsilon) = a(\pi)\varepsilon^{b(\pi)} + o(\varepsilon^{b(\pi)})$ and $q(\pi, \varepsilon) = \Theta(\varepsilon^{c(\pi)})$ for some integers $b(\pi)$ and $c(\pi)$, and a real number $a(\pi) > 0$. Then the kth relative moment can be written as

$$m_k(\varepsilon) = \sum_{\pi \in \Pi_B} q(\pi, \varepsilon) \left[\frac{p(\pi, \varepsilon)}{q(\pi, \varepsilon)\gamma(\varepsilon)} \right]^k$$

and the contribution of path $\pi \in \Pi_B$ to $m_k(\varepsilon)$ is

$$q(\pi, \varepsilon) \left[\frac{p(\pi, \varepsilon)}{q(\pi, \varepsilon)\gamma(\varepsilon)} \right]^k = \varepsilon^{\delta(k,\pi)} + o(\varepsilon^{\delta(k,\pi)}),$$

where $\delta(k, \pi) = k(b(\pi) - 6) - (k - 1)c(\pi)$. This contribution vanishes when $\varepsilon \to 0$ if and only if $\delta(k, \pi) > 0$. For the most likely path π_1, we have $\delta(k, \pi_1) = -(k - 1)c(\pi_1) \leq 0$ and its contribution to $m_k(\varepsilon)$ is $1+o(1)$ if and only if $q(\pi, \varepsilon) = \varepsilon^6 + o(\varepsilon^6)$. These two conditions are necessary and sufficient for having $m_k(\varepsilon) = 1 + o(1)$, i.e., for VRCM-$k$. To prove it formally, we actually have one more detail to check: the number of paths that contain cycles is infinite and we must make sure that their total contribution remains negligible. This is done for the general case in the next example. Note that in the present case, all cycles have probability $O(\varepsilon^2)$, so the probability of having c cycles or more decreases as $O(\varepsilon^{2c})$. Similarly, BRM-k holds if and only if $\delta(k, \pi) \geq 0$ for all acyclic paths $\pi \in \Pi_B$.

Table II enumerates all acyclic paths $\pi \in \Pi_B$, and gives the values of $b(\pi)$, $c(\pi)$,

and $\delta(k,\pi)$ for $k = 2$, 3, and 4, for those paths. We can see that VRCM-k holds for all $k < 3$ but not for $k \geq 3$. The problem comes from the path $\pi = ((0,0),(0,1),B)$, whose probability has not been increased sufficiently by the IS scheme. When this path is selected, the likelihood ratio is $\varepsilon^2/(1-\varepsilon^2)$, which decreases too slowly relative to the mean $\gamma(\varepsilon)$ when $\varepsilon \to 0$. The contribution of this path to the relative kth centered moment is

$$\Theta(\varepsilon^{12}|\varepsilon^2 - \varepsilon^6)/\varepsilon^6|^k) = \Theta(\varepsilon^{12}|\varepsilon^{-4} - 1|^k) = \Theta(\varepsilon^{-4(k-3)}),$$

which does not vanish as $\varepsilon \to 0$ for $k \geq 3$. For $k > 3$, this contribution actually increases with ε, so the estimator is not even BRM-k for $k > 3$. For $k = 3$, this contribution is $\Theta(1)$.

To improve this IS estimator and make it VRCM-k for all k, it suffices to change $q((0,0),(0,1),\varepsilon)$, say from ε^{12} to ε^8. Then, $c(\pi)$ decreases by 4 for the first three paths in Table II, and we have $\delta(k,\pi) > 0$ for all paths $\pi \in \Pi_B \setminus \{\pi_1\}$ and all k. The resulting estimator is VRCM-k for all k. We can also observe that changing from ε^{12} to ε^8 gives a better approximation of the zero-variance IS.

Example 2.23. We now develop the ideas of the previous example in a more general Markov chain setting. Consider a Markov chain $\{X_j,\, j \geq 0\}$ with finite state space and with transition probabilities

$$p(x,y,\varepsilon) = \mathbb{P}[X_j = y \mid X_{j-1} = x] = a(x,y)\varepsilon^{b(x,y)},$$

where $a(x,y)$ and $b(x,y)$ are non-negative constants (independent of ε) for all pairs of states (x,y). Let B be a given set of states and suppose that the chain starts from some fixed state $x_0 \notin B$. We want to estimate the probability $\gamma(\varepsilon)$ of reaching B before returning to x_0.

Let Π_B be the set of all sample paths $\pi = (x_0, x_1, \ldots, x_\tau)$ going from x_0 to B, where $x_\tau \in B$ and $x_j \notin B$ for all $j < \tau$. Suppose that among all the paths $\pi \in \Pi_B$, there is a set Π_1 of paths π having probability

$$p(\pi,\varepsilon) = \prod_{j=1}^{\tau} p(x_{j-1}, x_j, \varepsilon) = a(\pi)\varepsilon^b + o(\varepsilon^b)$$

where $a(\pi) > 0$ and $b > 0$, and all other paths have probability $p(\pi,\varepsilon) = o(\varepsilon^b)$. Suppose also that all cycles (paths going from one state to the same state) that belong to some path $\pi \in \Pi_B$ have probability $O(\varepsilon^\delta)$, for some constant $\delta > 0$. Then, Π_1 cannot contain paths having a cycle, so it must be finite. It is easy to see that the paths $\pi \in \Pi_1$ are the *dominant paths* within Π_B when $\varepsilon \to 0$, in the sense that

$$\lim_{\varepsilon\to 0} \frac{1}{\gamma(\varepsilon)} \sum_{\pi\in\Pi_1} p(\pi,\varepsilon) = \lim_{\varepsilon\to 0} \frac{a\varepsilon^b + o(\varepsilon^b)}{\gamma(\varepsilon)} = 1,$$

where $a = \sum_{\pi\in\Pi_1} a(\pi)$.

Suppose now that we simulate this chain using importance sampling by replacing the probabilities $p(x,y,\varepsilon)$ by new probabilities $q(x,y,\varepsilon)$ such that for any path

$\pi \in \Pi_1$, the new probability of that path satisfies

$$q(\pi,\varepsilon) = \prod_{j=1}^{\tau} q(x_{j-1}, x_j, \varepsilon) = \frac{a(\pi)}{a} + o(1) \tag{14}$$

when $\varepsilon \to 0$. This implies that the sum of probabilities of all paths in $\Pi_B \setminus \Pi_1$ is $o(1)$ under these new probabilities. The IS estimator of $\gamma(\varepsilon)$ is the likelihood ratio $Y(\varepsilon) = p(\pi,\varepsilon)/q(\pi,\varepsilon)$ if we reach B via some path π, and 0 if we do not reach B. When we reach B via a path $\pi \in \Pi_1$, we have

$$Y(\varepsilon) = p(\pi,\varepsilon)/q(\pi,\varepsilon) = \frac{a(\pi)\varepsilon^b}{a(\pi)/a + o(1)} = a\varepsilon^b + o(\varepsilon^b),$$

and this happens with probability $1 + o(1)$. The set of all other paths leading to B has total probability $o(1)$. We nevertheless need to bound the contribution of those paths to the moments of order $k > 1$, and this is a bit tricky because these paths could contain an unlimited number of cycles, so their number is generally infinite.

To bound the contribution of those paths $\pi \in \Pi_B \setminus \Pi_1$, we assume that for each such path having original probability $p(\pi,\varepsilon) = \Theta(\varepsilon^{b(\pi)})$ for $b(\pi) > b$, the new probability satisfies $q(\pi,\varepsilon) = \Theta(\varepsilon^{c(\pi)})$, for some constant $c(\pi) > 0$, and that these constants satisfy

$$\delta(k,\pi) = k[b(\pi) - b] - (k-1)c(\pi) > 0 \tag{15}$$

if we are interested in the kth moment. Finally, we assume that for any state $x \neq x_0$, $x \notin B$, and that belongs to a path $\pi \in \Pi_B$, the probability of returning to x (i.e., making a cycle) before hitting B or x_0 is never equal to 1 under the new probabilities, and the likelihood ratio associated with any such cycle does not exceed 1, at least for ε small enough. Since the number of possible cycles is finite, this assumption implies that there is a constant $\rho < 1$ such that the probability that there are j cycles or more does not exceed ρ^j. Let $\Pi_B^{(0)}$ be the set of paths in Π_B that contain no cycle. For any path $\pi \in \Pi_B$ that has cycles, let $\phi(\pi) \in \Pi_B^{(0)}$ the path obtained from π by removing all cycles. Under our assumptions, given that we have a path π for which $\phi(\pi) = \pi_0 \in \Pi_B^{(0)}$, the probability that this path has j cycles does not exceed ρ^j. Therefore, the set $\phi^{-1}(\pi_0)$ of all paths π that map to π_0 has total probability at most $q(\pi_0,\varepsilon)(1 + \rho + \rho^2 + \cdots) = q(\pi_0,\varepsilon)/(1-\rho)$. And the likelihood ratio associated with any path in $\phi^{-1}(\pi_0)$ does not exceed that of π_0 (for ε small enough). For the paths π for which $\pi_0 = \phi(\pi) \in \Pi_1$, the probability of a cycle must be $o(1)$, because $q(\pi,\varepsilon) = \Theta(1)$ if and only if $\pi \in \Pi_1$. We can then replace ρ by $o(1)$ in the above and the set of paths in $\phi^{-1}(\pi_0)$ that contain at least one cycle has total probability $q(\pi,\varepsilon)o(1)/(1-o(1))$.

With these ingredients in hand, we can bound the kth relative centered moment of the IS estimator as follows:

$$\mathbb{E}\left[\left|\frac{Y(\varepsilon)}{\gamma(\varepsilon)} - 1\right|^k\right]$$
$$= \sum_{\pi\in\Pi_B} q(\pi,\varepsilon)\left|\frac{p(\pi,\varepsilon)}{q(\pi,\varepsilon)\gamma(\varepsilon)} - 1\right|^k$$

$$
\begin{aligned}
&\leq \sum_{\pi\in\Pi_1} q(\pi,\varepsilon)\left|\frac{a\varepsilon^b+o(\varepsilon^b)}{\gamma(\varepsilon)}-1\right|^k + \sum_{\pi\in\Pi_1}\frac{q(\pi,\varepsilon)o(1)}{1-o(1)}\left|\frac{p(\pi,\varepsilon)}{q(\pi,\varepsilon)\gamma(\varepsilon)}-1\right|^k \\
&\quad + \sum_{\pi\in\Pi_B^{(0)}\setminus\Pi_1}\frac{q(\pi,\varepsilon)}{1-\rho}\left|\frac{p(\pi,\varepsilon)}{q(\pi,\varepsilon)\gamma(\varepsilon)}-1\right|^k \\
&= (1+o(1))\,|1+o(1)-1|^k + \sum_{\pi\in\Pi_1} o(1) + \sum_{\pi\in\Pi_B^{(0)}\setminus\Pi_1} O\left(\varepsilon^{c(\pi)}+\varepsilon^{k[b(\pi)-b]-(k-1)c(\pi)}\right) \\
&= o(1)
\end{aligned}
$$

when $\varepsilon \to 0$. So we have VRCM-k. From Proposition 2.21, this implies that $q(\pi,\varepsilon) - q^*(\pi,\varepsilon) \to 0$ for any sample path π, where $q^*(\pi,\varepsilon)$ denote the path probabilities under the zero-variance IS.

Condition (14) turns out to be also necessary for VRCM-k, since if $q(\pi,\varepsilon) = a(\pi)/a + \delta(\pi) + o(1)$ for some $\delta(\pi) \neq 0$ and $\pi \in \Pi_1$, then $Y(\varepsilon) = a(\pi)\varepsilon^b/[a(\pi)/a + \delta(\pi)+o(1)] = a\varepsilon^b/[1+a\delta(\pi)/a(\pi)]+o(\varepsilon^b)$, and the contribution of this path to the kth relative centered moment is no longer $o(1)$.

Example 2.22 does satisfy all the assumptions made here.

In the following sections, we examine the robustness concepts discussed so far in some settings that fit under the umbrella of estimating a first-passage probability for a Markov chain.

3. ESTIMATORS BASED ON ZERO-VARIANCE APPROXIMATION FOR FIRST-PASSAGE PROBABILITIES IN A MARKOV CHAIN

In this section, we adopt a framework where a rare event occurs when some discrete-time Markov chain hits a given set of states B before hitting some other set A, and we want to estimate the probability of that rare event. In some of these settings, the Markov chain is a random walk on the real line, with i.i.d. increments, and the rare event occurs when the walks exceeds some fixed level. We look at situations where the increments have light-tail and heavy-tail distributions, and we consider both state-independent and state-dependent IS schemes. Our purpose is to study, in these settings, the different robustness properties defined earlier, and to illustrate the differences between these properties.

The model is a Markov chain $X = \{X_j,\, j \geq 0\}$ living on a state space $\mathcal{S}$ equipped with a sigma-field $\mathcal{F}$, with transition kernel $K = \{K(x,C) : x \in S,\, C \in \mathcal{F}\}$. We use the notation $\mathbb{P}_x(\cdot)$ for the probability measure generated by X given that $X_0 = x$. For $C \subset \mathcal{S}$, define $\tau_C = \inf\{j \geq 0 : X_j \in C\}$. Given A and B, two disjoint subsets of $\mathcal{S}$, and some fixed initial state $x_0 \in (A \cup B)^c \stackrel{\text{def}}{=} \mathcal{S} \setminus A \cup B$, we are interested in estimating $\gamma(x_0)$, where

$$\gamma(x) = \gamma(x,\varepsilon) = \mathbb{P}_x[\tau_B < \tau_A]$$

is the probability of reaching B before A (in finite time) when starting from $x \in \mathcal{S}$. (We implicitly assume all along that $\tau_B < \tau_A$ implies that $\tau_B < \infty$.) In particular, $\gamma(x) = 1$ for $x \in B$ and $\gamma(x) = 0$ for $x \in A$. In this model, K, A, and B may depend on ε.

An importance sampling scheme here consists in replacing the kernel K by another kernel, and multiplying the original estimator by the appropriate likelihood ratio [Glynn and Iglehart 1989; Juneja and Shahabuddin 2006]. It is well-known that in this setting, a kernel K^* defined by

$$K^*(x, dy) = K(x, dy)\,\frac{\gamma(y)}{\gamma(x)}$$

for all x such that $\gamma(x) > 0$, and (say) $K^*(x, A) = 1$ when $\gamma(x) = 0$, gives a zero-variance IS estimator [Juneja and Shahabuddin 2006]. This kernel K^* describes the conditional behavior of the chain given the event $\{\tau_B < \tau_A\}$; see Blanchet and Glynn [2007], Theorem 1. Unfortunately, one cannot use it in practice to simulate the chain (in general), because this would require perfect knowledge of the function $\gamma(\cdot)$. But in view of this characterization of the optimal change-of-measure, a natural strategy in developing a state-dependent importance sampling for estimating $\gamma(x_0)$ is to use as a change-of-measure a transition kernel of the form

$$K_v(x, dy) = K(x, dy)\,\frac{v(y)}{w(x)},$$

where $v : \mathcal{S} \to [0, \infty)$ is a good approximation (in some sense) of the function $\gamma(\cdot)$, and

$$w(x) = \int_{\mathcal{S}} K(x, dy)\, v(y)$$

is the appropriate normalizing constant to make sure that $K_v(x, \cdot)$ integrates to 1. This $w(x)$ is assumed to be finite for every $x \in (A \cup B)^c$. We shall use $\mathbb{P}_x^v(\cdot)$ to denote the probability measure generated by the chain X under the kernel $K_v(\cdot)$, with initial state x, and $\mathbb{E}_x^v(\cdot)$ for the corresponding expectation. The corresponding IS estimator of $\gamma(x_0)$ is the indicator of the event multiplied by the likelihood ratio associated with the change of measure and the realized sample path:

$$Y = Y(\varepsilon) = \mathbb{I}[\tau_B < \tau_A] \prod_{j=1}^{\tau_B} \frac{w(X_{j-1})}{v(X_j)} = \mathbb{I}[\tau_B < \tau_A]\,\frac{v(X_0)}{v(X_{\tau_B})} \prod_{j=0}^{\tau_B - 1} \frac{w(X_j)}{v(X_j)}. \tag{16}$$

Since we know that $\gamma(x) = 1$ for $x \in B$, we can take $v(x) = 1$ for all $x \in B$. Note that when $v = \gamma$, we have $w = v$ and the last product in (16) equals 1. Ideally, we want v to be a good enough approximation to γ for this product to always remain close to 1; in that case Y will always take a value close to $\gamma(x_0)$ when $\tau_B < \tau_A$, which implies that most of the time the event $\{\tau_B < \tau_A\}$ will occur. Then, the variance of Y will be very small.

To rigorously prove robustness properties such as LE-k, BRM-k, and VRCM-k, we may use an asymptotic lower bound on $\gamma(x_0, \varepsilon)$ and an asymptotic upper bound on the kth moment of Y under the measure $\mathbb{P}_{x_0}^v(\cdot)$, for $\varepsilon \to 0$. The lower bound may come from a known asymptotic approximation of $\gamma(x_0, \varepsilon)$, while the upper bound can be obtained via a Lyapunov inequality as indicated in Proposition 3.1. This proposition generalizes a result of Blanchet and Glynn [2007], that corresponds to the case of $k = 2$, and which the authors have used to establish the BRE property of a state-dependent estimator.

PROPOSITION 3.1. *Suppose that there are two positive finite constants κ_1 and κ_2 and a function $h_k : \mathcal{S} \to [0, \infty)$ such that $v(x) \geq \kappa_1$ and $h_k(x) \geq \kappa_2$ for each $x \in B$, and*

$$\left(\frac{w(x)}{v(x)}\right)^k \mathbb{E}_x^v \left[h_k(X_1)\right] \leq h_k(x) \tag{17}$$

for all $x \in (A \cup B)^c$. Then, for all $x \in (A \cup B)^c$,

$$\mathbb{E}_x^v[Y^k] \leq \frac{v^k(x)h_k(x)}{\kappa_1^k \kappa_2}. \tag{18}$$

PROOF. Let $M = \{M_n,\, n \geq 0\}$ be defined via

$$M_n = h_k\left(X_{\tau_B \wedge n}\right) \prod_{j=0}^{\tau_B \wedge (n-1)} \left(\frac{w\left(X_j\right)}{v\left(X_j\right)}\right)^k \mathbb{I}\left(\tau_B \wedge n < \tau_A\right),$$

where $a \wedge b$ means $\min(a, b)$. We first show that under $\mathbb{P}_x^v(\cdot)$, M is a non-negative supermartingale adapted to the filtration $\mathcal{G} = \{\mathcal{G}_n = \sigma\left(X_0, ..., X_n\right),\, n \geq 0\}$ generated by the chain X. Let $\tau = \min(\tau_A, \tau_B) = \tau_{A \cup B}$ and note that τ is a stopping time with respect to $\mathcal{G}$, i.e., $\{\tau > n\} \in \mathcal{G}_n$ for all n.

We decompose

$$\mathbb{E}_x^v[M_{n+1} \mid \mathcal{G}_n] = \mathbb{E}_x^v\left[M_{n+1} \cdot \mathbb{I}(\tau > n) \mid \mathcal{G}_n\right] + \mathbb{E}_x^v\left[M_{n+1} \cdot \mathbb{I}(\tau \leq n) \mid \mathcal{G}_n\right]$$

and bound each of the two terms. We have

$$\begin{aligned}
&\mathbb{E}_x^v\left[M_{n+1} \cdot \mathbb{I}(\tau > n) \mid \mathcal{G}_n\right] \\
&\quad = \mathbb{I}\left(\tau > n,\, \tau_B \wedge n < \tau_A\right) \prod_{j=0}^{n-1} \left(\frac{w\left(X_j\right)}{v\left(X_j\right)}\right)^k \cdot \mathbb{E}_x^v\left[h_k\left(X_{n+1}\right) \left(\frac{w\left(X_n\right)}{v\left(X_n\right)}\right)^k \middle| \mathcal{G}_n\right] \\
&\quad \leq \mathbb{I}\left(\tau > n,\, \tau_B \wedge n < \tau_A\right) h_k\left(X_n\right) \prod_{j=0}^{n-1} \left(\frac{w\left(X_j\right)}{v\left(X_j\right)}\right)^k,
\end{aligned}$$

where the last inequality follows from (17). On the other hand,

$$\mathbb{E}_{x_0}^v\left[M_{n+1} \cdot \mathbb{I}\left(\tau \leq n\right) \mid \mathcal{G}_n\right] = h_k\left(X_{\tau_B}\right) \prod_{j=0}^{\tau_B - 1} \left(\frac{w\left(X_j\right)}{v\left(X_j\right)}\right)^k \mathbb{I}\left(\tau_B < \tau_A,\, \tau \leq n\right).$$

Combining these two inequalities, we obtain

$$\mathbb{E}_{x_0}^v\left[M_{n+1} \mid \mathcal{G}_n\right] \leq M_n.$$

It then follows from the supermartingale convergence theorem that

$$\lim_{n \to \infty} M_n = h_k\left(X_{\tau_B}\right) \prod_{j=0}^{\tau_B - 1} \left(\frac{w\left(X_j\right)}{v\left(X_j\right)}\right)^k \mathbb{I}\left(\tau_B < \tau_A\right)$$

almost surely. The supermartingale property further implies that

$$\mathbb{E}_{x_0}^v\left[M_n\right] \leq M_0 = h_k\left(x\right).$$

Fatou's lemma and the fact that $h_k(x) \geq \kappa_2$ for $x \in B$ imply that

$$\kappa_2 \mathbb{E}_x^v \left[\prod_{j=0}^{\tau_B - 1} \left(\frac{w(X_j)}{v(X_j)} \right)^k \mathbb{I}(\tau_B < \tau_A) \right] \leq h_k(x).$$

From this, we obtain that

$$\mathbb{E}_x^v \left[Y^k \right] = \mathbb{E}_x^v \left[\mathbb{I}[\tau_B < \tau_A] \left(\frac{v(x)}{v(X_{\tau_B})} \prod_{j=0}^{\tau_B - 1} \frac{w(X_j)}{v(X_j)} \right)^k \right] \leq \left(\frac{v(x)}{\kappa_1} \right)^k \frac{h_k(x)}{\kappa_2},$$

which yields the result. □

As a consequence of the previous proposition, we obtain

THEOREM 3.2. *Assume that the conditions of Proposition 3.1 are satisfied.*

(i) If

$$\lim_{\varepsilon \to 0} \frac{\ln[v(x_0, \varepsilon)] + k^{-1} \ln[h_k(x_0, \varepsilon)]}{\ln[\gamma(x_0, \varepsilon)]} = 1,$$

then $Y(\varepsilon)$ is LE-k.

(ii) If

$$\lim_{\varepsilon \to 0} \left[\frac{v(x_0, \varepsilon)}{\gamma(x_0, \varepsilon)} \right]^k h_k(x_0, \varepsilon) < \infty,$$

then $Y(\varepsilon)$ is BRM-k.

(iii) If

$$\lim_{\varepsilon \to 0} \left[\frac{v(x_0, \varepsilon)}{\gamma(x_0, \varepsilon)} \right]^k \frac{h_k(x_0, \varepsilon)}{\kappa_1^k \kappa_2} = 1,$$

then $Y(\varepsilon)$ is VRCM-k.

PROOF. The three assertions follow immediately from the corresponding definitions; for (iii), we use the equivalence given in Proposition 2.19. □

These sufficient conditions are often convenient to verify the BRM-k, LE-k, and VRCM-k properties of a given estimator. We will use the first two in the next section. It is clear that condition (iii) is much stronger than (ii), which is in turn stronger that (i). Dupuis and Wang [2004] have a similar condition for LE-2, and they interpret the Lyapunov function h_k as a *subsolution* to the recurrence equation of a stochastic game in which we select a change of measure (for IS) and then a devil picks a set of sample paths with the worst possible variance contribution.

4. LARGE DEVIATION PROBABILITIES IN RANDOM WALKS

4.1 The Random Walk

Let $D_1, D_2, \ldots$ be i.i.d. random variables, $S_j = D_1 + \cdots + D_j$ (the jth partial sum), for $j \geq 0$. Note that $\{S_j,\, j \geq 0\}$ is a random walk over the real line. Take a constant $\ell > \mathbb{E}[D_j]$, put $n = n(\varepsilon) = \lceil 1/\varepsilon \rceil$, and let

$$\gamma(\varepsilon) = \gamma(\varepsilon, \ell) = \mathbb{P}[S_n/n \geq \ell].$$

The weak law of large numbers guarantees that $\gamma(\varepsilon) \to 0$ when $\varepsilon \to 0$. The indicator function $Y(\varepsilon) = \mathbb{I}[S_n \geq n\ell]$ is an unbiased estimator of $\gamma(\varepsilon)$ with kth moment $\mathbb{E}[Y^k(\varepsilon)] = \gamma(\varepsilon)$, so its relative kth moment is

$$\gamma(\varepsilon)/\gamma^k(\varepsilon) = 1/\gamma^{k-1}(\varepsilon)$$

for all $k \geq 1$. Thus, this estimator is not LE-k whenever $k > 1$.

4.2 State-Independent Exponential Twisting Based on Large Deviation Theory

For this situation, it is well known that an LE-2 estimator can be obtained via IS with exponential twisting, under the assumption that D_j has a *light tail distribution* [Siegmund 1976; Bucklew et al. 1990; Bucklew 2004], as we now outline.

Suppose D_j has density π over $\mathbb{R}$, with finite *moment generating function*

$$M(\theta) = \int_{-\infty}^{\infty} e^{\theta x}\pi(x)dx = \mathbb{E}\left[e^{\theta D_j}\right]$$

for θ in a neighborhood of 0 (this is equivalent to assuming that D_j has finite moments of all orders). Let $\Psi(\theta) = \ln M(\theta)$ denote the *cumulant generating function*. Exponential twisting means inflating the density $\pi(x)$ by a factor that increases exponentially with x, and normalizing so that the new density integrates to 1. This new density is

$$\pi_\theta(x) = e^{\theta x}\pi(x)/M(\theta) = e^{\theta x - \Psi(\theta)}\pi(x), \qquad x \in \mathbb{R},$$

where $\theta > 0$ is a parameter to be determined and $M(\theta)$ turns out to be the appropriate normalization constant. Let $\mathbb{E}_\theta$ denote the mathematical expectation associated with the new density π_θ. It is easily seen that $\mathbb{E}_\theta[D_j] = \Psi'(\theta) = M'(\theta)/M(\theta)$ and $\Psi'(0) = M'(0) = \mu$. The IS estimator of $\gamma(\varepsilon)$ under this density is

$$Y(\theta, \varepsilon) = \mathbb{I}[S_n \geq n\ell]L(\theta, S_n),$$

where

$$L(\theta, S_n) = \exp[-\theta S_n]\ M^n(\theta) = \exp[n\Phi(\theta) - \theta S_n].$$

We now *assume* that there exists a real number $\theta^*_\ell > 0$ such that $\Psi'(\theta^*_\ell) = \ell$. This is typically the case because frequently, $\Psi'(\theta)$ is continuous in θ, $\Psi'(\theta) \to \infty$ when $\theta \to \theta_0$ for some $\theta_0 > 0$ (i.e. $\Psi'(\theta)$ is what is called *steep*) and we know that $\Psi'(0) = \mu < \ell$. Under *steepness*, the three propositions that follow are direct consequences of the results of Sadowsky [1993]. They imply that for all $k \geq 2$, $Y(\theta^*_\ell, \varepsilon)$ is LE-k but is not BRM-k. Sadowsky states his results only for integer k, but his proofs work for any real $k > 1$. Let $I(\ell) = \ell\theta^*_\ell - \Psi(\theta^*_\ell)$; this function I is known as the *large deviation rate function*.

PROPOSITION 4.1. *For any $k > 1$ and any θ the estimator $Y(\theta, \varepsilon)$ is not BRM-k. It is LE-k if and only if $\theta = \theta^*_\ell$. In the latter case,*

$$\lim_{\varepsilon\to 0} \frac{\ln \gamma(\varepsilon)}{n(\varepsilon)} = \lim_{\varepsilon\to 0} \frac{\ln \mathbb{E}[Y^k(\varepsilon)]}{kn(\varepsilon)} = I(\ell).$$

Suppose now that we make $m(\varepsilon)$ i.i.d. copies of $Y(\theta, \varepsilon)$, take their average $\tilde{\mu}(\varepsilon)$ as an estimator of μ and take their sample variance $\tilde{\sigma}^2(\varepsilon)$ as an estimator of the variance of $Y(\theta, \varepsilon)$.

PROPOSITION 4.2. *Suppose that $m(\varepsilon) \equiv m$ (a fixed constant). Then, for any $k \geq 1$, $\tilde{\sigma}^2(\varepsilon)$ is not BRM-k, and it is LE-k if and only if $\theta = \theta^*_\ell$.*

PROPOSITION 4.3. *Suppose that $\theta = \theta^*_\ell$. Then, for all $k > 1$, $\tilde{\mu}(\varepsilon)$ is BRM-k if and only if $m(\varepsilon) = \underline{O}(\varepsilon^{-1/2})$, and similarly for $\tilde{\sigma}^2(\varepsilon)$. On the other hand, these estimators have a computational cost proportional to $m(\varepsilon)n(\varepsilon) = \underline{O}(\varepsilon^{-3/2})$. Since their relative moments are $\Theta(1)$ when $m(\varepsilon) = \Theta(\varepsilon^{-1/2})$, their work-normalized relative variance is unbounded.*

4.3 A State-Dependent IS Scheme for Light-tailed Sums

BRM-k for $k > 1$ cannot be obtained with a state-independent IS scheme as in the previous section, but it can be achieved with a *state-dependent* IS scheme, as we now explain. As a key ingredient, we use the following (asymptotic) approximation of $\gamma(\varepsilon, \ell) = \mathbb{P}[S_n \geq n\ell]$, taken from Asmussen [2003], page 355:

PROPOSITION 4.4. *Assume that D_1 has a density with respect to the Lebesgue measure. Then, for fixed ℓ and $n \to \infty$,*

$$\mathbb{P}[S_n \geq n\ell] = \frac{\exp[-nI(\ell)]}{[2\pi n\Psi''(\theta^*_\ell)]^{1/2}\theta^*_\ell}[1 + o(1)]. \tag{19}$$

The random walk model considered here fits the framework of Section 3 if we define the state of the Markov chain at step j as $X_j = (n - j, L_j)$ where $n - j$ is the number of steps that remain and $(n - j)L_j = n\ell - S_j$ is the distance that remains to be covered for S_n to reach $n\ell$. We start in state $x_0 = (0, 0)$, the set B is $\{(0, \ell_n) : \ell_n \leq 0\}$, and we have

$$\gamma(n - j, \ell_j) = \mathbb{P}[S_n - S_j \geq \ell - (n - j)\ell_j] = \mathbb{P}[S_{n-j}/(n - j) \geq \ell_j].$$

In view of (19), we can think of approximating $\gamma(n - j, \ell_j)$ by

$$v(n - j, \ell_j) = \frac{\exp[-(n - j)I(\ell_j)]}{[2\pi(n - j)\Psi''(\theta^*_{\ell_j})]^{1/2}} \tag{20}$$

for $j < n$ and $\ell_j > 0$, and $v(0, \ell_n) = \mathbb{I}[\ell_n \leq 0]$. The latter ensures that we hit B with probability 1 under this IS scheme, because the last transition is made under the distribution conditional on hitting B. When $\ell_j \leq 0$ for $j < n$, IS is turned off for step j. For $j < n - 1$ and $x = (n - j, \ell_j)$ with $\ell_j > 0$, the normalizing constant $w(n - j, \ell_j)$ is

$$w(n - j, \ell_j) = \mathbb{E}^v_x \left[\frac{\exp[-(n - j - 1)I(L_{j+1})]}{[2\pi(n - j - 1)\Psi''(\theta^*_{L_{j+1}})]^{1/2}} \mid L_j = \ell_j \right]$$

where $L_{j+1} = [(n - j)L_j - D_{j+1})]/(n - j - 1)$. For $j = n - 1$, it is $w(1, \ell_{n-1}) = \mathbb{P}[D_n > \ell_{n-1}] = \gamma(1, \ell_{n-1})$. We have

$$\max_{n>j}[v(n - j, \ell_j)/\gamma(n - j, \ell_j)] < \infty$$

for any fixed j and ℓ_j. In our expression for v, we dropped the θ^*_ℓ that appears in the denominator of (19) because it typically leads to a simpler density and does not play a key role for the BRM-k property. If D_1 has the normal distribution, for

example, then the IS scheme without the θ_ℓ^* in the denominator just changes the parameters of the normal distribution.

Under the assumption that D_1 has the normal distribution, it is shown by Blanchet and Glynn [2006] that $w(n-j,\ell_j)/v(n-j,\ell_j) \le 1+(n-j)^{-2}$ for all $j < n$. In that case, to establish the BRM-k property, we can define

$$h_k(n-j,\ell_j) = \prod_{i=1}^{n-j}(1+i^{-2})^k$$

for $j \le n$, where an empty product equals 1 by convention. Then,

$$\begin{aligned}&\left(\frac{w(n-j,\ell_j)}{v(n-j,\ell_j)}\right)^k \mathbb{E}_x^v\left[\frac{h_k(n-j-1,L_{j+1})}{h_k(n-j,L_j)} \mid L_j=\ell_j\right]\\ =\ &\left(\frac{w(n-j,\ell_j)}{v(n-j,\ell_j)}\right)^k (1+(n-j)^{-2})^{-k} \ \le\ 1,\end{aligned}$$

so the conditions of Proposition 3.1 are satisfied with $\kappa_1 = \kappa_2 = 1$. Since the function h_k is bounded by $K = \prod_{i=1}^{\infty}(1+i^{-2})^k < \infty$, the BRM-$k$ property for all $k \ge 1$ then follows from Part (ii) of Theorem 3.2; this gives the following generalization of a result proved by Blanchet and Glynn [2006] for $k = 2$:

PROPOSITION 4.5. *Suppose that D_1 has a normal distribution. Then the IS scheme that approximates the zero-variance estimator as in Section 3 by using the function v defined in (20) as described above has the BRM-k property.*

Under the change-of-measure adopted for the previous result, the Gaussian property is preserved. That is, if the original (nominal) distribution of the $D_i's$ is a standard normal, then, given $S_k = s$ for $k < n-1$, D_{k+1} is normally distributed with mean $(n\ell - s)/(n-k-1)$ and variance $1+1/(n-k-1)$. This explicit description indicates why the estimator enjoys BRM-k. In particular, the twisting of the increment's mean is adjusted at each time-step to direct the process in the right direction and is turned off as the boundary $n\ell$ is approached. Although the variance is twisted incrementally, it is the contribution of the drift that drives the overshoot over the boundary in the standard (blind, or open loop) i.i.d. exponential tilting. In fact, in Blanchet et al. [2008], it is shown that it is possible to achieve BRM-k by tilting the mean only (not the variance), so the tilting applied to the variance, although convenient for the analysis because it comes from the asymptotic approximation (19), is not crucial. The zero-variance change-of-measure can be shown to yield an overshoot that remains bounded (in distribution) as $n \to \infty$ [Blanchet and Glynn 2006]. In contrast, because of the CLT, under the blind i.i.d. tilting the overshoot is of order $O(n^{1/2})$. Under the state-dependent importance sampling discussed here, the growth of the overshoot is controlled and its contribution when computing relative moments is well behaved. To get VRCM-k via Part (iii) of Theorem 3.2, we would need $h_k(x_0) = 1+o(1)$, which is not the case here. In fact, most of the contribution to the kth relative moment comes from the last few steps of the walk, and this contribution remains bounded away from 0 when $n \to \infty$.

A similar development can be made for the non-Gaussian case, where D_1 has a general distribution with finite moment generating function [Blanchet et al. 2008]. In fact, it turns out that BRM-k can be obtained by exponential twisting alone

if the twisting parameter is recomputed at each step. This is usually easier to implement than the zero-variance approximation based on (19).

4.4 A Criterion for Multidimensional Random Walks

Dupuis and Wang [2004] have developed a criterion that allows to design state-dependent IS estimators that are LE, in the context of a d-dimensional random walk with light-tailed increments. They restrict their change of measure to exponential twisting, but allow the twisting parameter to depend on the current state of the walk. The techniques can be extended to cover more general Markov processes [Dupuis and Wang 2005]. Here we summarize their results and argue that the resulting estimators are LE-k for all $k \geq 1$. Let $S_j = D_1 + \cdots + D_j$, where the D_j's are i.i.d. random variables with mean zero, taking their values in $\mathbb{R}^d$, and with cumulant generating function $\Psi(\theta) = \ln \mathbb{E}[\exp(\theta \cdot D_1)]$ for $\theta \in \mathbb{R}^d$. For simplicity, we assume that $\Psi(\cdot)$ is finite throughout $\mathbb{R}^d$.

We are interested in estimating $\mathbb{P}_0(S_n/n \in B)$, for a set $B \subset \mathbb{R}^d$ that does not contain 0. We assume as in Dupuis and Wang [2004] that the Legendre transform of Ψ, $L(\beta) = \sup_{\theta \in \mathbb{R}^d}(\theta \cdot \beta - \Psi(\theta))$ satisfies

$$\inf_{\beta \in \mathring{B}} L(\beta) = \inf_{\beta \in B} L(\beta) = \inf_{\beta \in \bar{B}} L(\beta)$$

where $\mathring{B}$ and $\bar{B}$ are the interior and closure of B, respectively. Note that the one-dimensional setting of Sections 4.1 and 4.2 is a special case of this with $B = [\ell, \infty)$; things are generally more complicated in the multidimensional case because we can reach B from many possible directions, whence the parameter β. We further assume that it is possible to find a function

$$I = \left\{ I(x,t) : x \in \mathbb{R}^d, 0 \leq t \leq 1 \right\},$$

that solves (in the classical sense) the nonlinear partial differential equation (PDE)

$$\partial_t I(x,t) = \Psi(-\nabla_x I(x,t)) \tag{21}$$

subject to $I(x,1) = 0$ for $x \in B$ and $I(x,1) = \infty$ for $x \notin B$. The algorithm suggested by Dupuis and Wang [2004] proceeds as follows. Let $x = S_j/n$ for some $j < n$; then let $t = j/n$ and define

$$\theta(x,t) = -\nabla_x I(x,t).$$

Sample the increment D_{j+1} according to the *twisted* distribution $\mathbb{P}_{\theta(x,t)}$ defined via

$$\mathbb{P}_{\theta(x,t)}(D_{j+1} \in dy) = \mathbb{P}(D_{j+1} \in dy) \exp[\theta(x,t)\, y - \Psi(\theta(x,t))].$$

The estimator takes the form

$$Y = \exp\left(\sum_{j=0}^{n-1} [-\theta(S_j, j/n)\, D_{j+1} + \Psi(\theta(S_j, j/n))] \right) \mathbb{I}(S_n/n \in B).$$

THEOREM 4.6. (Extends Dupuis and Wang [2004]). *Suppose that (21), with the boundary conditions given above, has a solution I in the classical sense. Let $\mathbb{P}_0^*(\cdot)$ be the probability measure generated by the previous state-dependent IS strategy, given*

$S_0 = 0$. *Then,*

$$\lim_{n\to\infty} -\frac{1}{n}\ln \mathbb{P}_0\left(S_n/n \in B\right) = \lim_{n\to\infty} -\frac{1}{nk}\ln \mathbb{E}_0^*[Y^k] = I(0,0),$$

so this estimator is LE-k for any $k \geq 1$.

Proof. We have written the description of the algorithm and the characterization of the solution to the Bellman equation derived by Dupuis and Wang [2004] in a slightly different way. Our description corresponds basically to the PDE approach derived in Section "Further remarks" of Dupuis and Wang [2004], pages 495–496. The Isaacs equation displayed on their page 495 can be solved and yields (21), which corresponds exactly to their equation (4.5), with our function I being denoted U in that paper. The proof that the algorithm verifies LE-k follows the same sequence of arguments as the proof of their Theorem 3.1 for LE-2, assuming that the Isaacs equation is satisfied in a classical sense. This equation holds if the solution to (21) is satisfied in the classical sense. The modifications to the proof are as follows (in their notation). Replace 2 by k in the definition of their function V^n, in the theorem's statement, and everywhere in the proof, including in the exponential that replaces the indicator in the middle of their page 490. We also multiply $-\langle\alpha, y\rangle + H(\alpha)$ and the function L by $k-1$ wherever they appear from the last line of page 490 up to Equation (3.8). To obtain the modification of (3.6), we apply their Lemma 7.1 with $f(y) = nW_F^n(x + y/n,\, i+1) + (k-2)[\langle\alpha, y\rangle - H(\alpha)]$. □

The previous result indicates that state-dependent samplers based on the solution of the Isaacs equation, proposed in Dupuis and Wang [2004] to design estimators that are asymptotically optimal, also achieve LE-k for $k > 2$. However, as pointed out in Section 3 of Dupuis and Wang [2004], in typical circumstances it is difficult (or impossible) to find a classical solution to the PDE (21). However, one often can introduce a mollification procedure, applied to a solution of this PDE in the weak sense (i.e., a solution for which the gradients are not strictly defined at every single point in time and space). Examples of such implementation schemes are described by Dupuis and Wang [2004] and also, in the case of a path-dependent simulation example, by Blanchet et al. [2006], both for $k = 2$. Similar techniques could be used for $k > 2$.

4.5 Heavy-Tailed Increments

We revisit the estimator proposed by Blanchet and Glynn [2007] for the steady-state delay in a single-server queue, and show that it can be designed to achieve BRM-k for all $k \geq 1$. The model is again a random walk over the real line.

We have $X_j = x_0 + D_1 + \cdots + D_j$ where the D_j's are i.i.d. with mean $\mathbb{E}[D_j] < 0$, and x_0 is some fixed constant. Let $B = B(\varepsilon) = [1/\varepsilon, \infty)$ and $A = \{\infty\}$, so $\tau_B = \inf\{j \geq 1 : X_j > 1/\varepsilon\}$ and $\tau_A = \infty$. We want to estimate $\gamma(\varepsilon) = \gamma(0, \varepsilon)$, where

$$\gamma(x, \varepsilon) = \mathbb{P}_x[\tau_B < \infty]$$

and $\mathbb{P}_x$ represents the probability when $x_0 = x$. This $\gamma(x, \varepsilon)$ may represent the probability of eventual ruin of an insurance company with initial reserve $-x + 1/\varepsilon$, using an appropriate interpretation of the D_j's in terms of i.i.d. claim sizes and

inter-arrival times. It can also be interpreted as the tail of the steady-state delay in a single-server queue [Asmussen 2003, page 260]. This model has other applications as well. Note that when $\varepsilon x < 1$, we have

$$\gamma(x,\varepsilon) = \gamma(0, \varepsilon/(1-\varepsilon x)). \tag{22}$$

To keep the discussion simple, we shall assume that D_j possesses a regularly varying tail; that is, for each $b > 0$,

$$\lim_{t\to\infty} \frac{P\left(D_j > bt\right)}{P\left(D_j > t\right)} = b^{-\alpha}$$

for some $\alpha > 1$. The discussion that follows holds in greater generality, for instance including Weibull or lognormal tails; see Blanchet and Glynn [2007], Section 3, for a more general framework.

In Blanchet and Glynn [2007], the authors propose to approximate $\gamma(\cdot)$ in the zero-variance change of measure by some function $v(\cdot)$ such that $\lim_{\varepsilon\to 0} v(x,\varepsilon)/\gamma(x,\varepsilon) = 1$, and suggest a specific selection of $v(\cdot)$ that is later proved to yield an IS estimator with BRE. More specifically, they introduce a non-negative random variable Z such that

$$\mathbb{P}[Z > t] = \min\left(1,\, \frac{1}{\mathbb{E}[-D_j]} \int_t^\infty \mathbb{P}[D_j > s] ds\right). \tag{23}$$

Motivated by a classical result stating that

$$\lim_{\varepsilon\to 0} \frac{\mathbb{P}[Z > 1/\varepsilon]}{\gamma(0,\varepsilon)} = 1, \tag{24}$$

(see, Asmussen [2003], page 296), and based on the discussion leading to Eq. (16), Blanchet and Glynn [2007] suggest using

$$v(x,\varepsilon) = v_{a^*}(x,\varepsilon) = \mathbb{P}[Z > a^* + 1/\varepsilon - x],$$

with corresponding normalization constant

$$w(x,\varepsilon) = w_{a^*}(x,\varepsilon) = \mathbb{P}[Z + D_j > a^* + 1/\varepsilon - x],$$

for some constant $a^* > 0$ chosen to satisfy the Lyapunov inequality of Proposition 3.1 for $k = 2$.

As we now show, for each $k \geq 1$, it is possible (and not difficult) to find a constant $a_k^* > 0$ that can be proved to yield the BRM-k property via Proposition 3.1. For this, we will use the following result, which follows directly from Proposition 3 of Blanchet and Glynn [2007].

PROPOSITION 4.7. *For each $k > 1$ and $\delta \in (0,1)$, there is a real number $a_k^* > 0$ such that*

$$-\delta \leq \frac{v_{a_k^*}^k(x,\varepsilon) - w_{a_k^*}^k(x,\varepsilon)}{\mathbb{P}[D_j > x + a_k^*] w_{a_k^*}^{k-1}(x,\varepsilon)} \tag{25}$$

for all $x \leq 1/\varepsilon$.

The constant a_k^* can be computed numerically, and the pair (δ, a_k^*) could eventually be selected to minimize the upper bound on the relative moment of order k given by the next theorem. This upper bound implies the BRM-k property.

THEOREM 4.8. *Fix $\delta \in (0,1)$, select $a_k^* > 0$ that satisfies (25), and let $\kappa(a_k^*) = \inf_{x \in B} v_{a_k^*}(x,\varepsilon) = \mathbb{P}[Z > a_k^*]$. Then,*

$$\mathbb{E}_x^v[Y^k] \leq \frac{v_{a_k^*}^k(x,\varepsilon)}{(1-\delta)(\kappa(a_k^*))^k}$$

and consequently

$$\limsup_{\varepsilon \to 0} \frac{\mathbb{E}_0^v\left[Y^k\right]}{\gamma^k(0,\varepsilon)} \leq \frac{1}{(1-\delta)(\kappa(a_k^*))^k} < \infty.$$

PROOF. Define

$$h_k(x) = \mathbb{I}(x - a_k^* \leq 1/\varepsilon) + (1-\delta)\mathbb{I}(x - a_k^* > 1/\varepsilon).$$

For $x \leq 1/\varepsilon$, the Lyapunov condition in Proposition 3.1 takes the form

$$\left(\frac{w_{a_k^*}(x,\varepsilon)}{v_{a_k^*}(x,\varepsilon)}\right)^{k-1} \frac{\mathbb{E}\left[v_{a_k^*}(D_1+x,\varepsilon)\, h_k(D_1+x)\right]}{v_{a_k^*}(x,\varepsilon)} \leq 1.$$

This is equivalent to

$$\frac{\mathbb{E}\left[v_{a_k^*}(D_1+x,\varepsilon)\, h_k(D_1+x)\right]}{w_{a_k^*}(x,\varepsilon)} \leq \left(\frac{w_{a_k^*}(x,\varepsilon)}{v_{a_k^*}(x,\varepsilon)}\right)^k. \tag{26}$$

Using the interpretation of $v_{a_k^*}(\cdot,\varepsilon)$ as a tail probability, we have

$$\begin{aligned}
&\frac{\mathbb{E}\left[v_{a_k^*}(D_1+x,\varepsilon)\,(h_k(D_1+x)-1)\right]}{w_{a_k^*}(x,\varepsilon)} \\
&= -\delta\, \frac{\mathbb{E}\left[\mathbb{P}(Z+D_1 > a_k^* + 1/\varepsilon - x \mid D_1) \cdot \mathbb{I}(D_1 > a_k^* + 1/\varepsilon - x)\right]}{\mathbb{P}(Z+D_1 > a_k^* + 1/\varepsilon - x)} \\
&= -\delta\, \mathbb{P}(D_1 > a_k^* + 1/\varepsilon - x \mid Z + D_1 > a_k^* + 1/\varepsilon - x).
\end{aligned}$$

Therefore, showing (26) is equivalent to establishing that

$$-\delta\, \mathbb{P}(D_1 > a_k^* + 1/\varepsilon - x \mid Z + D_1 > a_k^* + 1/\varepsilon - x) \leq \frac{v_{a_k^*}^k(x,\varepsilon) - w_{a_k^*}^k(x,\varepsilon)}{w_{a_k^*}^k(x,\varepsilon)}.$$

Since $Z \geq 0$, this in turn is equivalent to the inequality

$$-\delta \leq \frac{v_{a_k^*}^k(x,\varepsilon) - w_{a_k^*}^k(x,\varepsilon)}{P(D_1 > a_k^* + 1/\varepsilon - x)\, w_{a_k^*}^{k-1}(x,\varepsilon)},$$

which holds by definition of a_k^*. The conclusion then follows directly from Proposition 3.1 and the fact that $\lim_{\varepsilon \to 0} v(0,\varepsilon)/\gamma(0,\varepsilon) = 1$. □

The next example underlines the fact that finding an approximation v that provides BRM-k is not so obvious, and that the approximation must be good over a very wide range of states. In particular, it shows that even if $v(x,\varepsilon) = \gamma(x,\varepsilon)$ whenever ε is small enough, for any given x, one can still obtain an estimator that fails to achieve BRM-k or LE-k.

EXAMPLE 4.9. Suppose we take

$$v(x,\varepsilon) = \gamma(x,\varepsilon)\mathbb{I}\left(x \le c_\varepsilon\right) + \mathbb{I}\left(x > c_\varepsilon\right).$$

This gives

$$\begin{aligned} v(x,\varepsilon) &\le \ \gamma(c_\varepsilon,\varepsilon)\mathbb{I}\left(x \le c_\varepsilon\right) + \mathbb{I}\left(x > c_\varepsilon\right); \\ w(x,\varepsilon) &\ge \ \mathbb{P}\left(D_1 + x > c_\varepsilon\right). \end{aligned}$$

We will choose c_ε as a function of ε so that $c_\varepsilon \to \infty$. Then, for any fixed x, $v(x,\varepsilon) = \gamma(x,\varepsilon)$ when ε is small enough, which means that the function $v(\cdot)$ converges pointwise to $\gamma(\cdot)$ when $\varepsilon \to 0$. A natural question is if such approximation would be enough for BRM-k? We are interested in the kth moment

$$\begin{aligned} \mathbb{E}_0^v\left[Y^k\right] &= \mathbb{E}_0^v\left[\left(\prod_{j=0}^{\tau_B-1}\frac{w(X_k,\varepsilon)}{v(X_k,\varepsilon)}\right)^k \mathbb{I}\left(\tau_B < \infty\right)\right] \\ &= \mathbb{E}_0\left[\left(\prod_{j=0}^{\tau_B-1}\frac{w(X_k,\varepsilon)}{v(X_k,\varepsilon)}\right)^{k-1}\frac{\mathbb{I}\left(\tau_B < \infty\right)}{v(0,\varepsilon)}\right]. \end{aligned}$$

Our bounds on $w(x,\varepsilon)$ and $v(x,\varepsilon)$ imply that

$$\frac{w(x,\varepsilon)}{v(x,\varepsilon)} \ \ge \ \frac{\mathbb{P}\left(D_1 + x > c_\varepsilon\right)}{\gamma(c_\varepsilon,\varepsilon)\mathbb{I}\left(x \le c_\varepsilon\right) + \mathbb{I}\left(x > c_\varepsilon\right)} \ \ge \ \frac{\mathbb{P}\left(D_1 + x > c_\varepsilon\right)}{\gamma(c_\varepsilon,\varepsilon)}\mathbb{I}\left(x \le c_\varepsilon\right).$$

Therefore,

$$\mathbb{E}_0\left[\left(\prod_{j=0}^{\tau_B-1}\frac{w(X_k,\varepsilon)}{v(X_k,\varepsilon)}\right)^{k-1}\frac{\mathbb{I}\left(\tau_B < \infty\right)}{v(0,\varepsilon)}\right] \ \ge \ \left(\frac{\mathbb{P}\left(D_1 > c_\varepsilon\right)}{\gamma(c_\varepsilon,\varepsilon)}\right)^{k-1}\frac{\mathbb{P}_0\left(\tau_B = 1\right)}{v(0,\varepsilon)}.$$

For simplicity, let us assume that D_1 has Pareto-type tails with index $\alpha > 1$. In particular, $\mathbb{P}\left(D_1 > t\right)t^\alpha \to c > 0$ as $t \to \infty$ and, because of (22), (23, and (24),

$$\begin{aligned} \gamma(c_\epsilon,\epsilon) \ &= \ \gamma(0,\epsilon/(1-\epsilon c_\epsilon)) \ = \ \Theta\left(\mathbb{P}[Z > (1-\epsilon c_\epsilon)/\epsilon]\right) \\ &= \ \Theta\left(\int_{(1-\epsilon c_\epsilon)/\epsilon}^{\infty} c t^{-\alpha} dt\right) \ = \ \Theta((1-\epsilon c_\epsilon)/\epsilon)^{1-\alpha} \ = \ \Theta\left(\epsilon^{\alpha-1}\right) \end{aligned}$$

as $\varepsilon \to 0$ whenever $\varepsilon c_\varepsilon = o(1)$ as $\varepsilon \to 0$. Suppose we take $c_\varepsilon = \varepsilon^{-\beta}$ for some $\beta \in (0,1)$. Then, the right hand side of the previous inequality is $\Theta\left(\varepsilon^{\beta\alpha(k-1)}/\varepsilon^{k(\alpha-1)}\right)$, which blows up for $\varepsilon \to 0$ whenever $\beta < k(\alpha-1)/(\alpha(k-1))$.

The problem here is the contribution of the likelihood ratio corresponding to the interval $(c_\varepsilon, 1/\varepsilon]$ in the state space, due to a bad approximation of the zero-variance importance sampler in that region of the state space. The contribution of the likelihood ratio corresponding to this bad approximation is captured, most importantly, by the normalizing constant $w(x,\varepsilon)$, which involves a first transition expectation. This expectation must account for the possibility that the process jumps to the bad region and this possibility is quantified and added to the likelihood ratio. The accumulation of all these contributions induces a poor behavior of the

overall importance sampling strategy by inflating the moments of the likelihood ratio. This problem could be cured by increasing c_ε at a faster speed.

5. HIGHLY RELIABLE MARKOVIAN SYSTEMS

5.1 The Model

We consider an HRMS with c types of components and n_i components of type i, for $i = 1, \ldots, c$. Each component is either in a failed state or an operational state. The *state of the system* is represented by a vector $x = (x^{(1)}, \ldots, x^{(c)})$, where $x^{(i)}$ is the number of *failed* components of type i. Thus, we have a finite state space $\mathcal{S}$ of cardinality $(n_1 + 1) \cdots (n_c + 1)$. We suppose that $\mathcal{S}$ is partitioned in two subsets $\mathcal{U}$ and $\mathcal{F}$, where $\mathcal{U}$ is a decreasing set (i.e., if $x \in \mathcal{U}$ and $x \geq y \in \mathcal{S}$, then $y \in \mathcal{U}$) that contains the state $\mathbf{0} = (0, \ldots, 0)$ in which all the components are operational. We say that $y \prec x$ when $y \leq x$ and $y \neq x$.

Following Shahabuddin [1994], we assume that the times to failure and times to repair of the individual components are independent exponential random variables with respective rates

$$\lambda_i(x) = a_i(x)\varepsilon^{b_i(x)} \quad \text{and} \quad \mu_i(x) = \Theta(1) \tag{27}$$

for type-i components when the current state is x, where $a_i(x) > 0$ and $b_i(x) \geq 1$ are real numbers for each i. The parameter $\varepsilon \ll 1$ represents the rarity of failures; the failure rates tend to zero when $\varepsilon \to 0$. The choice of parameterization determines in what asymptotic regime the system is studied. Failure propagation is allowed: from state x, there is a probability $p_i(x, y)$ (which may depend on ε) that the failure of a type-i component directly drives the system to state y, in which there could be additional component failures. Thus, the net jump rate from x to y is

$$\lambda(x, y) = \sum_{i=1}^{c} \lambda_i(x) p_i(x, y) = O(\varepsilon).$$

Similarly, the repair rate from state x to state y is $\mu(x, y)$ (with possible grouped repairs), where $\mu(x, y)$ does not depend on ε (i.e., repairs are not rare events when they are possible). The system starts in state $\mathbf{0}$ and we want to estimate the probability $\gamma(\varepsilon)$ that it reaches the set $\mathcal{F}$ before returning to state $\mathbf{0}$. Estimating this probability is relevant in many practical situations [Heidelberger 1995; Juneja and Shahabuddin 2006].

This model evolves as a continuous-time Markov chain (CTMC) $(Z(t),\, t \geq 0\}$, where $Z(t)$ is the system's state at time t. Its canonically embedded discrete time Markov chain (DTMC) is $\{X_j,\, j \geq 0\}$, defined by $X_j = Z(\xi_j)$ for $j = 0, 1, 2, \ldots$, where $\xi_0 = 0$ and $0 < \xi_1 < \xi_2 < \cdots$ are the jump times of the CTMC. Since the quantity of interest here, $\gamma(\varepsilon)$, does not depend on the jump times of the CTMC, it suffices to simulate the DTMC. This chain $\{X_j,\, j \geq 0\}$ has transition probability matrix $\mathbf{P}$ with elements

$$\mathbf{P}(x, y) = \mathbb{P}[X_j = y \mid X_{j-1} = x] = \lambda(x, y)/q(x)$$

if the transition from x to y corresponds to a failure and

$$\mathbf{P}(x, y) = \mu(x, y)/q(x)$$

if it corresponds to a repair, where

$$q(x) = \sum_{y \in \mathcal{S}} (\lambda(x,y) + \mu(x,y))$$

is the total jump rate out of x, for all x, y in $\mathcal{S}$. We will use $\mathbb{P}$ to denote the corresponding measure on the sample paths of the DTMC.

To fit the framework of Section 3, we must distinguish two cases for state $\mathbf{0}$: (1) when we are in the initial state $X_0 = \mathbf{0}$ and (2) if we return to that state later on. We consider them as two different states; in the second case, we will call the state $\mathbf{0}'$ to make the distinction. Then, we have $A = \{\mathbf{0}'\}$, $B = \mathcal{F}$, and $\gamma(\varepsilon) = \mathbb{P}[\tau_B < \tau_A]$.

Let Γ denote the set of pairs $(x, y) \in \mathcal{S}^2$ for which $\mathbf{P}(x, y) > 0$. Our final assumptions are that the DTMC is irreducible on $\mathcal{S}$ and that for every state $x \in \mathcal{S}$, $x \neq \mathbf{0}$, there exists a state $y \prec x$ such that $(x, y) \in \Gamma$ (that is, at least one repairman is active whenever a component is failed). We further assume that from state $\mathbf{0}$, the failures with probability in $\Theta(1)$ do not directly lead to $\mathcal{F}$, since otherwise $\gamma(\varepsilon) = \Theta(1)$ is not a rare event probability. Shahabuddin [1994] shows that for this model, there is a real number $r > 0$ such that $\gamma(\varepsilon) = \Theta(\varepsilon^r)$, i.e., the probability of interest decreases at a polynomial rate when $\varepsilon \to 0$. Nakayama [1996] makes the additional assumption that the $b_i(x)$ are positive integers; in that case, r is always an integer. We also make this assumption for the remainder of the paper, to simplify the analysis.

5.2 IS for the HRMS Model

Several IS schemes have been proposed in the literature for this HRMS model; see, e.g., Cancela et al. [2002], Nakayama [1996], Shahabuddin [1994]. Here we first limit ourselves to the so-called *simple failure biasing* (SFB), also named *Bias1*, and then consider more general classes of changes of measures determined by certain sets of conditions. Our aim is to analyze the robustness properties under that scheme, and not to try approaching the zero-variance IS as in Example 2.23. We do not claim that SFB is a good IS scheme. SFB changes the matrix $\mathbf{P}$ to a new matrix $\mathbf{P}^*$ defined as follows. For states $x \in \mathcal{F} \cup \{\mathbf{0}\} \cup \{\mathbf{0}'\}$, we have $\mathbf{P}^*(x, y) = \mathbf{P}(x, y)$ for all $y \in \mathcal{S}$, i.e., the transition probabilities are unchanged. For any other state x, a fixed probability ρ is assigned to the set of all failure transitions, and a probability $1 - \rho$ is assigned to the set of all repair transitions. In each of these two subsets, the individual probabilities are taken proportionally to the original ones. Note that the IS does not depend on the parameterization by ε; it depends only on the actual rates. Under certain additional assumptions, this change of measure increases the probability of failure when the system is up, in a way that failure transitions are no longer rare events, i.e., $\mathbb{P}^*[\tau_B < \tau_A] = \Theta(1)$.

For a given sample path ending at step $\tau = \min(\tau_A, \tau_B)$, the likelihood ratio for this change of measure can be written as

$$L = L(X_0, \ldots, X_\tau) = \frac{\mathbb{P}[(X_0, \ldots, X_\tau)]}{\mathbb{P}^*[(X_0, \ldots, X_\tau)]} = \prod_{j=1}^{\tau} \frac{\mathbf{P}(X_{j-1}, X_j)}{\mathbf{P}^*(X_{j-1}, X_j)}$$

and the corresponding (unbiased) IS estimator of $\gamma(\varepsilon)$ is given by

$$Y(\varepsilon) = L(X_0, \ldots, X_\tau)\, \mathbb{I}\,[\tau_B < \tau_A]\,. \tag{28}$$

We will now examine the robustness properties of this estimator under the SFB sampling.

5.3 Asymptotic Robustness for the HRMS Model Under IS

For this HRMS model, a characterization of the IS schemes that satisfy the BRE property was obtained by Nakayama [1996] and the equivalence between BRE and LE for this model was mentioned without proof in Heidelberger [1995]. Our first result generalizes this, for SFB. Note that under a static change of measure such as SFB, the expected computing time is $\Theta(1)$.

PROPOSITION 5.1. *In the HRMS framework adopted here, with SFB, the two properties BRM-k and LE-k are equivalent. These two properties are also equivalent for the g-th empirical moment of $Y(\varepsilon)$ and for its empirical variance.*

PROOF. Recall that Shahabuddin [1994] proves that $\gamma(\varepsilon) = \Theta(\varepsilon^r)$ for some integer $r \geq 0$. Following the same argument, just replacing the likelihood ratio L by L^g, we can show (as done by Tuffin [1999] for the second moment) that there is a constant $s_g \leq gr$ such that

$$\mathbb{E}[Y^g(\varepsilon)] = \Theta(\varepsilon^{s_g}), \tag{29}$$

where $Y(\varepsilon)$ is defined in (28). Note that $s_1 = r$. From Jensen's inequality, we also have $s_{kg} \leq k s_g$. The equivalence between LE-k and BRM-k for the g-th empirical moment then follows from Example 2.13. The case of the empirical variance is handled by replacing $Y(\varepsilon)$ by $S_n^2(\varepsilon)$; one can see that each moment of $S_n^2(\varepsilon)$ is $\Theta(\varepsilon^\nu)$ for some $\nu \geq 0$ and the result follows easily from that and Example 2.13. □

Our next result characterizes BRM-k for the g-th empirical moment in the HRMS framework. In particular, it gives characterizations of BRM-k for $Y(\varepsilon)$, as well as BRM-k and LE-k for $S_n^2(\varepsilon)$. It requires additional notation. We no longer limit our change of measure for IS to SFB, but we restrict it to a class $\mathcal{I}$ of measures $\mathbb{P}^*$ defined by a transition probability matrix $\mathbf{P}^*$ with the following property: whenever $(x, y) \in \Gamma$ and $\mathbf{P}(x, y) = \Theta(\varepsilon^d)$, then $\mathbf{P}^*(x, y) = \Theta(\varepsilon^\ell)$ for $\ell \leq d$. This means that the probability of a transition under the new probability transition matrix is never significantly smaller than under the original one. From now on, we assume that $\mathbf{P}^*$ satisfies this property. Note that SFB and all other IS schemes developed in the literature belong to this class.

We define the following sets of sample paths:

$$\begin{aligned}\Delta_m \;=\; \{&(x_0, \cdots, x_n) :\; n \geq 1, x_0 = \mathbf{0}, x_n \in \mathcal{F},\\ &x_j \notin \{\mathbf{0}', \mathcal{F}\} \text{ and } (x_{j-1}, x_j) \in \Gamma \text{ for } 1 \leq j \leq n,\\ &\text{and } \mathbb{P}\left[(X_0, \cdots, X_\tau) = (x_0, \cdots, x_n)\right] = \Theta(\varepsilon^m)\};\\ \Delta \;=\; &\bigcup_{m=r}^{\infty} \Delta_m;\end{aligned}$$

this Δ is the set of all paths that lead to the rare event.

We now derive a necessary and sufficient condition on $\mathbb{P}^*$ for BRM-k of the g-th moment. This result means that a path cannot be too rare under the IS measure

$\mathbb{P}^*$ to verify BRM-k for the g-th moment. Special cases of this result were obtained under the same conditions in Nakayama [1996] for BRE ($k = 2$ and $g = 1$), where it was shown that $\ell \le 2m - r$ is needed, and in Tuffin [1999] and [Tuffin 2004] for BNA, where the necessary and sufficient condition is $\ell \le 3m/2 - 3s/4$, where s is the real number such that $\sigma^2(\varepsilon) = \Theta(\varepsilon^s)$. Note that $s = s_2$ if and only if $\sigma^2(\varepsilon) = \Theta(Y^2(\varepsilon))$, where s_g is defined via (29).

THEOREM 5.2. *For an IS measure $\mathbb{P}^* \in \mathcal{I}$, we have BRM-k for the g-th empirical moment if and only if for all integers m such that $r \le m < ks_g$ and all paths $(x_0, \cdots, x_n) \in \Delta_m$,*

$$\mathbb{P}^*\{(X_0, \cdots, X_\tau) = (x_0, \cdots, x_n)\} = \Theta(\varepsilon^\ell)$$

for some $\ell \le k(mg - s_g)/(kg - 1)$.

PROOF. For $k = g = 1$, the interval for m is empty and we always have BRM-1 for the first moment, so the result holds. We now suppose that $kg > 1$.

(a) Necessary condition. Suppose that there exist $m \in \mathbb{N}$ and $(x_0, \cdots, x_n) \in \Delta_m$ such that $\mathbb{P}^*\{(X_0, \cdots, X_\tau) = (x_0, \cdots, x_n)\} = \Theta(\varepsilon^{k(mg-s_g)/(kg-1)+\ell'})$ with $\ell' > 0$ and $m < ks_g$. Then we have

$$\begin{aligned}\mathbb{E}[(Y(\varepsilon))^{kg}] &\ge L(x_0, \cdots, x_n)^{kg}\, \mathbb{P}^*\left[(X_0, \cdots, X_\tau) = (x_0, \cdots, x_n)\right] \\ &= \Theta(\varepsilon^{kg(m-k(mg-s_g)/(kg-1)-\ell')+k(mg-s_g)/(kg-1)+\ell'}) \\ &= \Theta(\varepsilon^{ks_g-(kg-1)\ell'}).\end{aligned}$$

Thus $\mathbb{E}[(Y(\varepsilon))^{kg}]/\mathbb{E}[(Y(\varepsilon))^g]^k = \underline{O}(\varepsilon^{-(kg-1)\ell'})$, which is unbounded when $\varepsilon \to 0$.

(b) Sufficient condition. Let $(x_0, \cdots, x_n) \in \Delta_m$ such that $m < ks_g$. Under the given condition, we have

$$\mathbb{P}^*\left[(X_0, \cdots, X_\tau) = (x_0, \cdots, x_n)\right] = \Theta(\varepsilon^\ell)$$

for some $\ell \le k(mg - s_g)/(kg - 1)$. Then,

$$(L(x_0, \cdots, x_n))^{kg}\, \mathbb{P}^*\left[(X_0, \cdots, X_\tau) = (x_0, \cdots, x_n)\right] = \frac{\Theta(\varepsilon^{kgm})}{\Theta(\varepsilon^{kg\ell})}\Theta(\varepsilon^\ell) = O(\varepsilon^{ks_g}).$$

Using the fact that $|\Delta_m| < \infty$ from the first part of Lemma 1 ii) of Nakayama [1996], we have

$$\sum_{r \le m < ks_g} \sum_{(x_0, \cdots, x_n) \in \Delta_m} (L(x_0, \cdots, x_n))^{kg}\, \mathbb{P}^*\left[(X_0, \cdots, X_{\tau_F}) = (x_0, \cdots, x_n)\right] = O(\varepsilon^{ks_g}). \tag{30}$$

Also, using again Lemma 1 of Nakayama [1996] (with N the total number of components, and α, β and δ constant),

$$\begin{aligned}&\sum_{m=ks_g}^{\infty} \sum_{(x_0, \cdots, x_n) \in \Delta_m} [L(x_0, \cdots, x_n)]^{kg}\ \mathbb{P}^*\left[(X_0, \cdots, X_\tau) = (x_0, \cdots, x_n)\right] \\ \le &\sum_{m=ks_g}^{\infty} \sum_{(x_0, \cdots, x_n) \in \Delta_m} \delta^{m+1}\alpha\beta^m\varepsilon^m\end{aligned}$$

$$\leq \sum_{m=ks_g}^{\infty} |\mathcal{S}|^{(m+1)N} \delta^{m+1} \alpha \beta^m \varepsilon^m$$
$$= \alpha\delta|\mathcal{S}|^N \sum_{m=ks_g}^{\infty} \left(|\mathcal{S}|^{(m+1)}\delta\beta\varepsilon\right)^m$$
$$= \Theta(\varepsilon^{ks_g}). \tag{31}$$

Combining (30) and (31) gives $\mathbb{E}[(Y(\varepsilon))^{kg}] = O(\varepsilon^{ks_g})$, meaning that we have BRM-k of the g-th moment. □

In Tuffin [1999], a different class $\mathcal{J}$ of measures $\mathbb{P}^*$ defined by a transition probability matrix $\mathbf{P}^*$ is used, motivated by the fact that absolute centered moments were considered. This class is more restrictive: for such a $\mathbf{P}^*$, whenever $(x, y) \in \Gamma$ and $\mathbf{P}(x, y) = \Theta(\varepsilon^d)$, if $y \succ x \neq \mathbf{0}$, then $\mathbf{P}^*(x, y) = \Theta(\varepsilon^\ell)$ with $\ell < d$, whereas if $x \succ y$ or if $y \succ x = \mathbf{0}$, then $\mathbf{P}^*(x, y) = \Theta(\varepsilon^\ell)$ with $\ell \leq d$. Using this class of measures, we could show, by similar arguments to those above and in Tuffin [1999] and Tuffin [2004], that we have BRM-k for the g-th moment if and only if for all integers ℓ and m such that $m - \ell < r$, and all $(x_0, \cdots, x_n) \in \Delta_m$ with $\mathbb{P}^*\{(X_0, \cdots, X_\tau) = (x_0, \cdots, x_n)\} = \Theta(\varepsilon^\ell)$, we have $\ell \leq k(mg - s_g)/(kg - 1)$. The difference in the characterization is then in terms of the set of paths. Note that the set here is more restrictive (because the class of functions is more restrictive too). Indeed, if $m - \ell < r = s_1$, then $m < \ell + s_1 \leq ks_g$ for all $g \geq 2$.

In the specific case of the empirical mean and variance, we have the following:

Corollary 5.3. *For an IS measure $\mathbb{P}^* \in \mathcal{I}$, we have BRM-k for $Y(\varepsilon)$ if and only if for all integers m such that $r \leq m < kr$ and all $(x_0, \cdots, x_n) \in \Delta_m$,*

$$\mathbb{P}^*\{(X_0, \cdots, X_\tau) = (x_0, \cdots, x_n)\} = \Theta(\varepsilon^\ell)$$

for $\ell \leq k(m - r)/(k - 1)$. We also have BRM-k for $Y^2(\varepsilon)$ if and only if the same condition holds with $\ell \leq k(2m - s_2)/(2k - 1)$. We have BRM-k for the empirical variance if and only if $\ell \leq k(2m - s)/(2k - 1)$.

The following additional relationships between measures of robustness were proved in Tuffin [2004]:

Proposition 5.4. *In our HRMS framework with an IS sampling scheme in $\mathcal{J}$, BNA implies AGEV, which implies BRE, which implies AGEM. For each of these implications, the converse is not true.*

The next result implies that IS sampling schemes in $\mathcal{I}$ cannot provide VRMC-k.

Proposition 5.5. *In our HRMS setting, with an IS measure in $\mathcal{I}$, we have $\mathbb{E}[(Y(\varepsilon) - \gamma(\varepsilon))^k] = \underline{O}(\gamma^k(\varepsilon))$ for all $k \geq 1$. In particular, $\sigma^2(\varepsilon) = \underline{O}(\gamma^2(\varepsilon))$.*

Proof. From our assumptions, there is a path $\pi = (\mathbf{0}, x, \ldots, \mathbf{0}')$ that does not hit $\mathcal{F}$, such that the initial failure leading to the transition from $\mathbf{0}$ to x has probability $\Theta(1)$ (because no repair is possible from state $\mathbf{0}$), and thereafter has only repairs until we return to $\mathbf{0}'$. (We must have $x \notin \mathcal{F}$ because otherwise $\gamma(\varepsilon) = \Theta(1)$.) This path has probability $\Theta(1)$ under IS. Since

$$\mathbb{E}[(Y(\varepsilon) - \gamma(\varepsilon))^k] = \mathbb{E}\left[(L\,\mathbb{I}(\tau_B < \tau_A) - \gamma(\varepsilon))^k\right]$$

$$\begin{aligned} &\geq \gamma^k(\varepsilon)\,\mathbb{P}\left[(X_0,\cdots,X_\tau)=\pi\right] \\ &= \Theta(\gamma^k(\varepsilon)), \end{aligned}$$

we get that $\mathbb{E}[(Y(\varepsilon)-\gamma(\varepsilon))^k] = \underline{O}(\gamma^k(\varepsilon))$. □

Our necessary and sufficient conditions in Theorem 5.2 lead to the following results.

PROPOSITION 5.6. *For an IS scheme in $\mathcal{I}$, BRM-k and LE-k for the g-th moment are equivalent. Similarly, for $S_n^2(\varepsilon)$, BRE and LE, are equivalent.*

PROOF. The first part follows again directly from Example 2.13, using the fact that $\mathbb{E}[(Y(\varepsilon))^g] = \Theta(\varepsilon^{s_g})$ and $\mathbb{E}[(Y(\varepsilon))^{kg}] = \Theta(\varepsilon^{s_{kg}})$ with $s_{kg} \leq k s_g$ from Jensen's inequality. For the empirical variance, we use the arguments of the same examples, combined with the fact that $\sigma^2(\varepsilon) = \Theta(\varepsilon^s)$ and $\mathbb{E}[S_n^4(\varepsilon)] = \Theta(\varepsilon^t)$ with $t \leq 2s$. □

Next we show that BRM-2 and LE-2 for $S_n^2(\varepsilon)$ are stronger than BNA when using the class of measures $\mathcal{J}$.

PROPOSITION 5.7. *Under an importance measure in $\mathcal{J}$, BRM-2 for $S_n^2(\varepsilon)$ implies BNA.*

PROOF. This is a direct consequence of the necessary and sufficient conditions over the paths for the BNA and BRM-2 properties. These conditions are that for all ℓ and m such that $m - \ell < r$, and such that there is a path $(x_0,\cdots,x_n) \in \Delta$ for which $\mathbb{P}\{(X_0,\cdots,X_\tau) = (x_0,\cdots,x_n)\} = \Theta(\varepsilon^m)$ and $\mathbb{P}^*\{(X_0,\cdots,X_\tau) = (x_0,\cdots,x_n)\} = \Theta(\varepsilon^\ell)$, we must have $\ell \leq 4m/3 - 2s/3$ for BRM-2 for $S_n^2(\varepsilon)$ and $\ell \leq 3m/2 - 3s/4$ for BNA. But $4m/3 - 2s/3 = 8/9(3m/2 - 3s/4)$, so the theorem is proved if we always have $3m/2 - 3s/4 \geq 0$, i.e., $2m \geq s$, which is true since $2m \geq 2r \geq s$. □

The following counter-example shows that the converse is not true: there are systems and IS measures $\mathbb{P}^*$ for which BNA is verified but not BRM-2 for $S_n^2(\varepsilon)$.

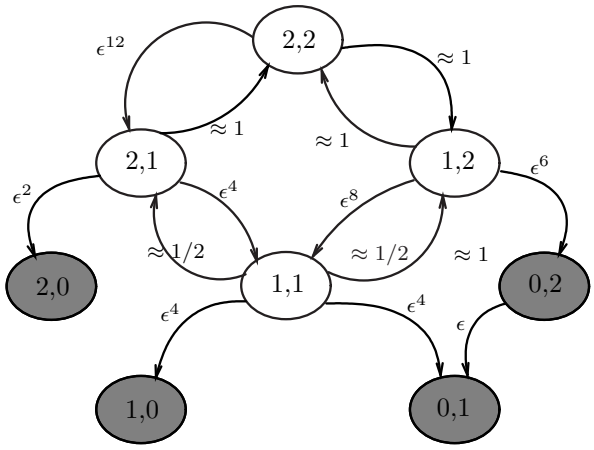

Fig. 1. A two-dimensional model with its transition probabilities.

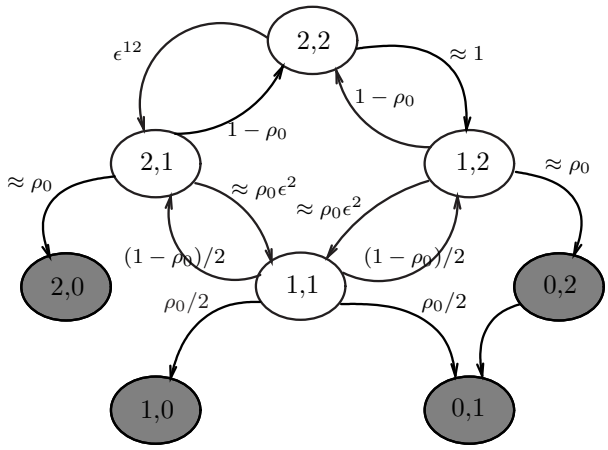

Fig. 2. A two-dimensional example with SFB transition probabilities.

EXAMPLE 5.8. We consider the same system as in Example 2.22, with two component types and two components of each type. The original transition probabilities are shown in Figure 1, and those using SFB failure biasing can be seen in Figure 2. The states in $\mathcal{F}$ are colored in gray. For this model, as can be easily seen in Figure 1, $r = 6$ and Δ_6 is comprised of the single path $((2,2),(1,2),(0,2))$. Moreover, $s = s_2 = 12$ and the sole path in Δ such that

$$\frac{\mathbb{P}^2\{(X_0,\cdots,X_\tau) = (x_0,\cdots,x_n)\}}{\mathbb{P}^*\{(X_0,\cdots,X_\tau) = (x_0,\cdots,x_n)\}} = \Theta(\varepsilon^{12})$$

is the path in Δ_6 for which Figure 2 shows that it is $\Theta(1)$ under probability measure $\mathbb{P}^*$. If ℓ is the integer such that $\mathbb{P}\{(X_0,\cdots,X_\tau) = (x_0,\cdots,x_n)\} = \Theta(\varepsilon^m)$ and $\mathbb{P}^*\{(X_0,\cdots,X_\tau) = (x_0,\cdots,x_n)\} = \Theta(\varepsilon^\ell)$, it can also be readily checked that $\ell \leq 3m/2 - 3s/4$ for all paths, meaning that BNA is verified. However, the path $((2,2),(2,1),(2,0))$ is such that $m = 14$ and $\ell = 12$. Then $12 = \ell > 4m/3 - 2s/3 = 32/3$, so the necessary and sufficient condition of Theorem 5.2 for $k = g = 2$ is not verified. So we have BNA but not BRM-2 for $S_n^2(\varepsilon)$. It is also easy to verify that for this example, we have BRM-3 but not BRM-4.

6. CONCLUSION

We have introduced and studied several new characterizations of the asymptotic robustness of estimators in the context of rare-event simulation. For $k > 2$, the new properties of BRM-k and LE-k are relevant whenever we estimate higher moments than the mean. The new concept of VRCM-k is much stronger than the standard concepts of BRE and LE, which have been the usual targets when defining IS schemes over the last decade. The design of estimators with the VRCM-k property is a quite interesting challenge for the coming years. For certain classes of applications, this could lead to much more efficient estimators than those currently available. In fact, such estimators have already started to appear very recently. Another important topic for further research is the development of an appropri-

ate framework to analyze work-normalized versions of the asymptotic robustness properties examined here. It would have to address the difficulties discussed in the introduction.

ACKNOWLEDGMENTS

This work has been supported by NSERC-Canada grant No. ODGP0110050 and a Canada Research Chair to the first author, NSF grant DMS 0595595 to the second author, EuroFGI Network of Excellence and INRIA's cooperative research initiative RARE to the third author, and the Binational Science Foundation Grant 2002284 to the fourth author. The first author thanks the IRISA for its generous support during his sabbatical in Rennes, where much of this paper was written. The comments of the Guest Editors, associate editor, and two reviewers have helped improving the paper.

REFERENCES

Asmussen, S. 2002. Large deviations in rare events simulation: Examples, counterexamples, and alternatives. In *Monte Carlo and Quasi-Monte Carlo Methods 2000*, K.-T. Fang, F. J. Hickernell, and H. Niederreiter, Eds. Springer-Verlag, Berlin, 1–9.

Asmussen, S. 2003. *Applied Probability and Queues*. Springer-Verlag, New York, NY.

Bentkus, V. and Götze, F. 1996. The Berry-Esseen bound for Student's statistic. *The Annals of Probability 24,* 1, 491–503.

Blanchet, J. and Glynn, P. W. 2006. Strongly efficient estimators for light-tailed sums. In *Proceedings of ValueTools 2006: International Conference on Performance Evaluation Methodologies and Tools*. ACM Publications, Pisa, Italy.

Blanchet, J. and Glynn, P. W. 2007. Efficient rare-event simulation for the maximum of a random walk with heavy-tailed increments. Manuscript.

Blanchet, J., Glynn, P. W., and Liu, J. C. 2006. State-dependent importance sampling and large deviations. In *Proceedings of the Sixth International Workshop on Rare Event Simulation*, W. Sandmann, Ed. Bamberg, Germany, 154–161.

Blanchet, J., Leder, K., and Glynn, P. W. 2008. Efficient simulation of light-tailed sums: an old folk song sung to a faster new tune. Manuscript.

Bolhuis, P. G., Chandler, D., Dellago, C., and Geissler, P. L. 2002. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annual Review of Physical Chemistry 53*, 291–318.

Boots, N. K. and Shahabuddin, P. 2000. Simulating $GI/GI/1$ queues and insurance processes with subexponential distributions. In *Proceedings of the 2000 Winter Simulation Conference*. IEEE Press, 656–665.

Bourin, C. and Bondon, P. 1998. Efficiency of high-order moment estimates. *IEEE Transactions on Signal Processing 46,* 1, 255–258.

Bucklew, J., Ney, P., and Sadowsky, J. S. 1990. Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains. *Journal of Applied Probability 27*, 44–59.

Bucklew, J. A. 2004. *Introduction to Rare Event Simulation*. Springer-Verlag, New York.

Cancela, H., Rubino, G., and Tuffin, B. 2002. MTTF estimation by Monte Carlo methods using Markov models. *Monte Carlo Methods and Applications 8,* 4, 312–341.

Cancela, H., Rubino, G., and Tuffin, B. 2005. New measures of robustness in rare event simulation. In *Proceedings of the 2005 Winter Simulation Conference*, M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, Eds. IEEE Press, 519–527.

Dupuis, P. and Wang, H. 2004. Importance sampling, large deviations, and differential games. *Stochastics and Stochastics Reports 76*, 481–508.

Dupuis, P. and Wang, H. 2005. Dynamic importance sampling for uniformly recurrent Markov chains. *Annals of Applied Probability 15*, 1–38.

Durrett, R. 1996. *Probability: Theory and Examples*, second ed. Duxbury Press.

Ermakov, S. M. and Melas, V. B. 1995. *Design and Analysis of Simulation Experiments.* Kluwer Academic, Dordrecht, The Netherlands.

Feller, W. 1971. *An Introduction to Probability Theory and Its Applications, Vol. 2*, second ed. Wiley, New York, NY.

Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering.* Springer-Verlag, New York.

Glasserman, P., Heidelberger, P., Shahabuddin, P., and Zajic, T. 1998. A large deviations perspective on the efficiency of multilevel splitting. *IEEE Transactions on Automatic Control AC-43,* 12, 1666–1679.

Glasserman, P., Heidelberger, P., Shahabuddin, P., and Zajic, T. 1999. Multilevel splitting for estimating rare event probabilities. *Operations Research 47,* 4, 585–600.

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

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

Goyal, A., Shahabuddin, P., Heidelberger, P., Nicola, V. F., and Glynn, P. W. 1992. A unified framework for simulating Markovian models of highly reliable systems. *IEEE Transactions on Computers C-41*, 36–51.

Hammersley, J. M. and Handscomb, D. C. 1964. *Monte Carlo Methods.* Methuen, London.

Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation 5,* 1, 43–85.

Juneja, S. 2007. Estimating tail probabilities of heavy tailed distributions with asymptotically zero relative error. *QUESTA 57*, 115–127.

Juneja, S. and Shahabuddin, P. 2006. Rare event simulation techniques: An introduction and recent advances. In *Simulation*, S. G. Henderson and B. L. Nelson, Eds. Handbooks in Operations Research and Management Science. Elsevier, Amsterdam, The Netherlands, 291–350. Chapter 11.

Kalos, M. H. and Whitlock, P. A. 1986. *Monte Carlo Methods.* John Wiley & Sons.

Katz, M. 1963. A note on the Berry-Esseen theorem. *Annals of Mathematical Statistics 34*, 1007–1008.

Kendall, M. and Stuart, A. 1977. *The advanced theory of statistics*, 4th ed. Vol. 1: Distribution theory. Macmillan, New York, NY.

L'Ecuyer, P., Demers, V., and Tuffin, B. 2007. Rare-events, splitting, and quasi-Monte Carlo. *ACM Transactions on Modeling and Computer Simulation 17,* 2, Article 9.

L'Ecuyer, P. and Tuffin, B. 2008a. Approximate zero-variance simulation. In *Proceedings of the 2008 Winter Simulation Conference.* IEEE Press, 170–181.

L'Ecuyer, P. and Tuffin, B. 2008b. Approximating zero-variance importance sampling in a reliability setting. *Annals of Operations Research*. Submitted.

Lewis, E. E. and Böhm, F. 1984. Monte Carlo simulation of Markov unreliability models. *Nuclear Engineering and Design 77*, 49–62.

Nakayama, M. K. 1996. General conditions for bounded relative error in simulations of highly reliable Markovian systems. *Advances in Applied Probability 28*, 687–727.

Petrov, V. V. 1995. *Limit Theorems in Probability Theory.* Oxford University Press, Oxford, U.K.

Sadowsky, J. S. 1993. On the optimality and stability of exponential twisting in Monte Carlo estimation. *IEEE Transactions on Information Theory IT-39*, 119–128.

Shahabuddin, P. 1994. Importance sampling for the simulation of highly reliable Markovian systems. *Management Science 40,* 3, 333–352.

Siegmund, D. 1976. Importance sampling in the Monte Carlo study of sequential tests. *The Annals of Statistics 4*, 673–684.

Tuffin, B. 1999. Bounded normal approximation in simulations of highly reliable Markovian systems. *Journal of Applied Probability 36,* 4, 974–986.

Tuffin, B. 2004. On numerical problems in simulations of highly reliable Markovian systems. In *Proceedings of the 1st International Conference on Quantitative Evaluation of SysTems (QEST)*. IEEE CS Press, University of Twente, Enschede, The Netherlands, 156–164.

Villén-Altamirano, M. and Villén-Altamirano, J. 2006. On the efficiency of RESTART for multidimensional systems. *ACM Transactions on Modeling and Computer Simulation 16,* 3, 251–279.

Wilks, S. S. 1962. *Mathematical Statistics*. Wiley, New York, NY.

Submitted August 30, 2007. Revised August 14, 2008; October 28, 2008; December 23, 2008.