4

Robustness properties and confidence interval reliability issues

Peter W. Glynn, Gerardo Rubino and Bruno Tuffin

4.1 Introduction

In this chapter, we discuss the robustness and reliability of the estimators of the probability of a rare event (or, more generally, of the expectation of some function of rare events) with respect to rarity: is the estimator accurate as rarity increases? (recall that accuracy, when estimating small probabilities, focuses on relative rather than absolute errors). And what about the reliability (i.e., the coverage) of the associated confidence interval?

If we parameterize the model with a (small) real ε such that the probability of the rare event considered decreases to zero as $\varepsilon \to 0$, we need to control the quality of the estimator as rarity increases, with respect to accuracy and coverage. An estimator will be said to be *robust* (in different senses defined hereafter) if its quality (i.e., the gap with respect to the true value) is not significantly affected when $\varepsilon \to 0$. Similarly, an estimator is always accompanied with a confidence interval. A *reliable* estimator is then an estimator for which the confidence interval coverage does not deteriorate as $\varepsilon \to 0$. Those two notions are different: one focuses on the error itself, the other on quality of the error estimation.

Rare Event Simulation using Monte Carlo Methods Edited by Gerardo Rubino and Bruno Tuffin
 ISBN: 978-0-470-77269-0

To better illustrate this, let us start with the standard or crude estimator of the probability of a rare event. Let ε be this probability and $(X_i)_{1\leq i\leq n}$ be independently and identically distributed random variables such that $X_i = 1$ if the rare event occurs at the ith trial and 0 otherwise. The standard estimator of ε is $\widehat{\gamma}_n^{\mathrm{STD}} = n^{-1}\sum_{i=1}^n X_i$. The sum $\sum_{i=1}^n X_i$ is a binomial random variable with variance $n\varepsilon(1-\varepsilon)$, and the resulting confidence interval for ε, centered at $\widehat{\gamma}_n^{\mathrm{STD}}$, at confidence level $1-\alpha$, is

$$\left[\widehat{\gamma}_n^{\mathrm{STD}} - z_{1-\alpha/2}\frac{\sqrt{\varepsilon(1-\varepsilon)}}{\sqrt{n}}, \widehat{\gamma}_n^{\mathrm{STD}} + z_{1-\alpha/2}\frac{\sqrt{\varepsilon(1-\varepsilon)}}{\sqrt{n}}\right]$$

where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and Φ is the standard normal cumulative distribution function. The relative half-width RE of the confidence interval is therefore $z_{1-\alpha/2}\sqrt{1-\varepsilon}/\sqrt{n\varepsilon}$. For a fixed sample size n, this means that, as $\varepsilon \to 0$, the relative error of the estimation goes to infinity. Therefore, the accuracy of the estimator deteriorates as $\varepsilon \to 0$. The *absolute* error given by the confidence interval half-width $z_{\alpha/2}\sqrt{\varepsilon(1-\varepsilon)}/\sqrt{n}$ tends to 0 with ε, but at the much smaller rate $\sqrt{\varepsilon}$ than ε, so it does not give a good idea of the order of magnitude of the probability of interest. In other words, in order to get a fixed relative half-width $\mathrm{RE} = \delta$ of the confidence interval as $\varepsilon \to 0$, one would have to increase the sample size (which usually means the simulation computating time) as

$$n = (z_{1-\alpha/2})^2\frac{1-\varepsilon}{\delta^2\varepsilon},$$

that is, in inverse proportion to ε. The aim of rare event simulation is to construct estimators for which the relative error is kept under control as the event probability decreases to zero. Such estimators are said to be *robust*, and families of *robusness properties* will be discussed in this chapter.

But looking only at the (theoretical) relative error, or some of its closely related notions introduced below, may be hazardous, or may only provide partial views of the possible problems. When evaluating γ using some unbiased estimator $\widehat{\gamma}_n = n^{-1}\sum_{i=1}^n X_i$, where the X_i are independently and identically (generally) distributed random variables with mean μ and variance σ^2, not only is $\mathbb{E}(\widehat{\gamma}_n) = \gamma$ unknown in practice, but so is its variance $\mathrm{Var}(\widehat{\gamma}_n) = \sigma_n^2 = \sigma^2/n$. Generally σ^2 is estimated by the unbiased $\widehat{\sigma}_n^2$:

$$\sigma^2 \approx \widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \widehat{\gamma}_n)^2.$$

This estimator is at least as sensitive to rarity as $\widehat{\gamma}_n$ itself.

Returning to the crude estimation of a probability ε by the average of Bernoulli random variables $\widehat{\gamma}_n^{\mathrm{STD}}$, if n is much smaller than $1/\varepsilon$, the rare event will most likely not be observed (on average, an occurrence appears after $1/\varepsilon$ replications), leading to a confidence interval $(0, 0)$ because $\widehat{\gamma}_n = \widehat{\sigma}_n^2 = 0$. With the (very unlikely) assumption that we end up with exactly one occurrence

of the rare event, $\widehat{\gamma}_n = 1/n$ overestimates the event, and the variance is also overestimated by $\widehat{\sigma}_n^2 = 1/n$. We then get a large confidence interval, with a very high coverage in this (very unlikely) case. This highlights not only the problem of robustness of the estimator, but also the problem of the *reliability*, meaning the error in terms of coverage, of the confidence interval produced. As stated before, the two notions are different: robustness is about the actual error with respect to the true value, while reliability is about the coverage of the confidence interval, both as the probability of the rare event goes to zero.

Note that for binomial random variables, such as the one we were looking at, we know how to generate a more reliable confidence interval even for small probabilities ε. For instance, the Wilson score interval gives an interval

$$\left(\frac{\widehat{\gamma}_n^{\mathrm{STD}} + \frac{1}{2n} z_{1-\alpha/2}^2 \pm z_{1-\alpha/2} \sqrt{\frac{\widehat{\gamma}_n^{\mathrm{STD}}(1-\widehat{\gamma}_n^{\mathrm{STD}})}{n} + \frac{z_{1-\alpha/2}^2}{4n^2}}}{1 + \frac{1}{n} z_{1-\alpha/2}^2} \right)$$

(but note that there exist other interval constructions; see [11] for a description and some comparisons). This interval is known to yield a better reliability, but is very conservative for fixed n as ε decreases. The relative half-width of the confidence interval, on the other hand, is still growing to infinity as ε tends to zero.

This chapter investigates the robustness properties and reliability issues in rare event simulation. Section 4.2 quickly reviews the known robustness properties in the literature, including bounded relative error (also called bounded relative variance), and logarithmic efficiency (also called asymptotic optimality). Section 4.3 discusses the efficiency of an estimator when computation time is taken into account. Section 4.4 discusses the related notion of reliability of the corresponding confidence interval. We start by illustrating in Section 4.4.1 that bad rare event estimations are not always checked by looking at intervals of the form $(0, 0)$, but can be much more difficult to detect. We then present two reliability measures. Section 4.5 summarizes the chapter by setting out some practical rules for detecting the presence of problems associated with the reliability of the observed confidence interval. Section 4.6 concludes the chapter.

4.2 Classical asymptotic robustness properties

This section describes the basic asymptotic robustness properties that can be found in the literature. For a recent survey, the reader is advised to look at [7], where more definitions are covered and discussed in detail.

As noted before, if we want to investigate the robustness properties of estimators with respect to rarity, it is very useful to parameterize the model. Let $\gamma = \gamma(\varepsilon)$ be the expectation (or probability if we restrict ourselves to integrating indicator functions) we are trying to estimate, parameterized by ε and such that $\gamma(\varepsilon) \to 0$ as $\varepsilon \to 0$. In this way the event can be arbitrarily small by playing with the value of ε, which allows the behavior of the estimator to be captured as rarity increases.

Consider an unbiased estimator $\widehat{\gamma}_n$ of γ, built from a sample having size n. The bounded relative error (BRE) is defined in [14]. It basically states that the relative half-width confidence interval already studied above is bounded uniformly in ε, for a fixed sample size n. This asserts that the relative error is not sensitive to the rarity of the event and is then the typical desirable property.

Definition 1. *Let σ_n^2 denote the variance of the estimator $\widehat{\gamma}_n$, $\sigma_n = \sqrt{\sigma_n^2}$ and let z_δ denote the $1 - \delta/2$ quantile of the standard normal distribution ($z_\delta = \Phi^{-1}(1 - \delta/2)$ where Φ is the standard normal cumulative distribution). Recall that the relative error* RE *associated with $\widehat{\gamma}_n$ is defined by the half-width confidence interval*

$$\mathrm{RE} = z_\delta \frac{\sigma_n}{\gamma}. \tag{4.1}$$

We say that we have a bounded relative error if RE *remains bounded as $\varepsilon \to 0$ (i.e., uniformly in ε).*

This property has been extensively studied and is often seen as the key property to verify [6, 8].

The aforementioned crude estimator is a typical illustration of one not verifying BRE. Additionally, increasing the occurrence of the rare event might not be sufficient. On the other hand, some estimators do possess the BRE property. Those two assertions are verified by the next two examples.

Consider the following example taken from [16], which can be seen as a simple case of the Markovian dependability models described in Chapter 6.

Example 1. A system consists of two types of components with two components of each type. Failure rates are $o(\varepsilon)$ for some parameter ε, and the transition probabilities of the embedded discrete-time Markov chain are as described in Figure 4.1, where (i, j) denotes the state with i (j) operational components of type 1 (2). The states where the system is down are shaded gray. We see that the system is functioning as soon as there is at least one component of each class that is operational.

Associated with each transition we put the first term of the development of the corresponding probability in powers of ε. We want to estimate the probability γ that, starting from $(2, 2)$, we reach a down state before returning to $(2, 2)$.

Given the target γ, we can simplify the model by collapsing or aggregating the failed states into a single one which we make absorbing. The resulting chain is shown in Figure 4.2.

Since $\gamma \ll 1$ because $\varepsilon \ll 1$ (we will see that $\gamma \approx 2\varepsilon^2$), we use the importance sampling (IS) method, and specifically the failure biasing scheme (see Section 6.3.2), with transition probabilities described in Figure 4.3. Basically, for each functioning state different from the initial $(2, 2)$, we increase the probability of failure to the constant q and use individual probabilities proportional to the original ones. The parameter q is chosen between $1/2$ and 1, for instance,

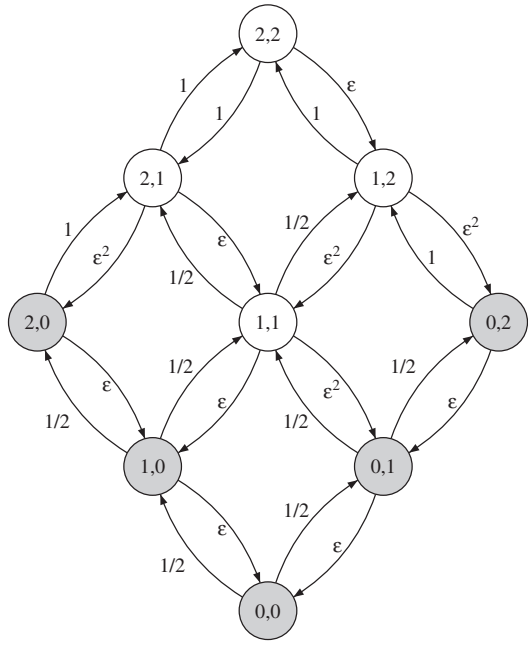

Figure 4.1 The evolution of a four-component system with two classes of components, subject to failures and repairs. The scheme shows the canonically embedded discrete-time Markov chain, where we give the simplest equivalents of the transition probabilities as $\varepsilon \to 0$.

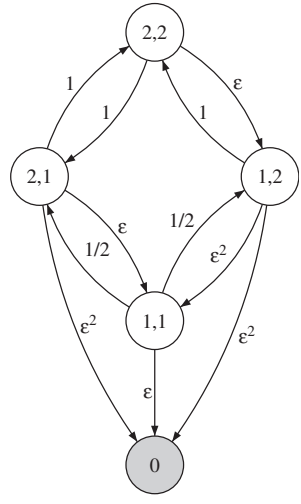

Figure 4.2 The result of aggregating the failed states in previous chain into a single absorbing one.

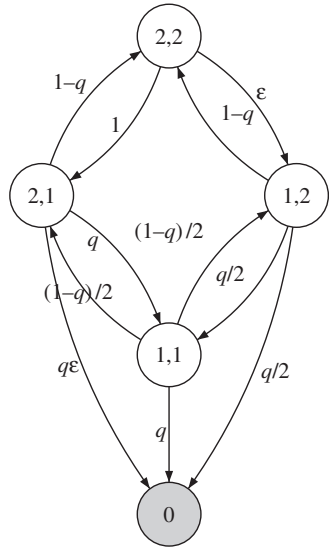

Figure 4.3 The result of changing the measure according to the failure biasing scheme with parameter q, again indicating the equivalents of the transition probabilities.

$q = 0.8$. The idea is, more generally, to enforce the transition probability associated with a failure to some $\Theta(1)$ value, instead of $o(1)$.

As seen in Chapter 1, the probability γ is given by

$$\gamma = \sum_{\pi \in \mathcal{P}_F} p(\pi),$$

where $\mathcal{P}_F$ is the set of all paths starting at (2,2), ending at a down state, and not visiting either (2,2) or a failed state in between, and $p(\pi)$ is the probability of path π under the original measure.

In this simple chain, there are six elementary paths in $\mathcal{P}_F$ (an elementary path is a path not visiting the same state more than once): $\pi_1 = ((2, 2), (2, 1), (0))$; $\pi_2 = ((2, 2), (2, 1), (1, 1), (0))$; $\pi_3 = ((2, 2), (2, 1), (1, 1), (1, 2), (0))'$ $\pi_4 = ((2, 2), (1, 2), (0))$; $\pi_5 = ((2, 2), (1, 2), (1, 1), (0))$; $\pi_6 = ((2, 2), (1, 2), (1, 1), (2, 1), (0))$. Their corresponding probabilities are $p(\pi_1) \approx \varepsilon^2$, $p(\pi_2) \approx \varepsilon^2$, $p(\pi_3) \approx \varepsilon^3/2$, $p(\pi_4) \approx \varepsilon^3$, $p(\pi_5) \approx \varepsilon^4$, $p(\pi_6) \approx \varepsilon^5/2$.

Observe that any other path include cycles that always strictly increase the order of the path probability in ε. This means that there are only a finite number of paths having the same order k in ε for any k, and thus, that $\gamma = 2\varepsilon^2 + o(\varepsilon^2)$ because of the two dominant paths π_1 and π_2 [14].

Let us now consider the IS scheme. To explore its performance, we must evaluate the variance of the IS estimator $\widehat{\gamma}_n^{\mathrm{IS}}$. For this purpose, denoting by Π a generic random path and by $\widetilde{p}(\pi)$ the probability of path π under the new

measure, we write

$$\mathrm{Var}(\widehat{\gamma}_n^{\mathrm{IS}}) = \frac{1}{n}\left\{\widetilde{\mathbb{E}}[L^2(\Pi)1(\Pi \in \mathcal{P}_F)] - \gamma^2\right\} = \frac{1}{n}\left[\sum_{\pi\in\mathcal{P}_F} \frac{p^2(\pi)}{\widetilde{p}(\pi)} - \gamma^2\right],$$

where $\widetilde{\mathbb{E}}$ denotes the expectation with respect to the IS measure. Looking at the probability of the six paths under the IS measure, the dominant term in this sum comes from π_1; it is in ε^3, and we get

$$\mathrm{Var}(\widehat{\gamma}_n^{\mathrm{IS}}) = \frac{\varepsilon^3}{nq} + o(\varepsilon^3).$$

The relative error of the IS estimator is $\mathrm{RE} = 1.96\sqrt{\mathrm{Var}(\widehat{\gamma}_n^{\mathrm{IS}})}/(\widehat{\gamma}_n^{\mathrm{IS}}\sqrt{n})$. We see that RE is proportional to $1/\sqrt{\varepsilon}$ and thus goes to infinity as $\varepsilon \to 0$.

Example 2. Consider a system failing according to an exponential distribution with rate λ. We wish to compute the probability γ that the system fails before ε. For such a trivial problem, we know that $\gamma = 1 - e^{-\lambda\varepsilon}$. Assume that we want to estimate this number using IS, and that we still sample from an exponential density, but with a different rate $\widetilde{\lambda}$. Our IS estimator is the random variable $X = 1_{[0,T]}L$ with L the likelihood ratio. The second moment of this estimator is

$$\widetilde{\mathbb{E}}[X^2] = \int_0^{\varepsilon} \left(\frac{\lambda e^{-\lambda y}}{\widetilde{\lambda} e^{-\widetilde{\lambda} y}}\right)^2 \widetilde{\lambda} e^{-\widetilde{\lambda} y} dy = \frac{\lambda^2}{\widetilde{\lambda}(2\lambda - \widetilde{\lambda})}(1 - e^{-(2\lambda-\widetilde{\lambda})\varepsilon}).$$

The relative error $z_\delta\sigma/\gamma$ is bounded if and only if $\widetilde{\mathbb{E}}[X^2]/\gamma^2$ is bounded as $\varepsilon \to 0$. It can easily be seen that, if $\widetilde{\lambda} = 1/\varepsilon$,

$$\frac{\widetilde{\mathbb{E}}[X^2]}{\gamma^2} = \frac{\lambda^2(1 - e^{-(2\lambda-\widetilde{\lambda})\varepsilon})}{\widetilde{\lambda}(2\lambda - \widetilde{\lambda})(1 - e^{-\lambda\varepsilon})^2} \longrightarrow e - 1 \quad \text{as } \varepsilon \to 0.$$

So, RE remains bounded as $\varepsilon \to 0$.

BRE has often been found difficult to verify in practice. For this reason, people often use logarithmic efficiency, also called asymptotic optimality.

Definition 2. *An unbiased estimator $\widehat{\gamma}_n$ of γ is said to be logarithmic efficient with respect to rarity parameter ε if*

$$\lim_{\varepsilon\to 0} \frac{\ln \mathbb{E}[\widehat{\gamma}_n^2]}{\ln \gamma} = 2.$$

Note that the quantity under limit is always positive and less than or equal to 2. This is because $\mathrm{Var}(\widehat{\gamma}_n) \geq 0$, so $\mathbb{E}[\widehat{\gamma}_n^2] \geq \gamma^2$ and then $\ln \mathbb{E}[\widehat{\gamma}_n^2] \geq 2\ln\gamma$.

Basically, this property means that the second moment and the square of the mean go to zero at the same *exponential* rate. Asymptotic optimality has been widely used in queuing applications, for the IS class of simulation methods (see Chapter 5).

It can be proved that asymptotic optimality is a necessary but not sufficient condition for BRE. Indeed, if the relative error corresponding to estimator $\widehat{\gamma}_n$ of γ is bounded, then there is some $\kappa > 0$ such that $E[\widehat{\gamma}^2] \leq \kappa^2\gamma^2$, that is, $\ln E[\widehat{\gamma}_n^2] \leq \ln\kappa^2 + 2\ln\gamma$, leading to $\lim_{\varepsilon\to 0} \ln E[\widehat{\gamma}_n^2]/\ln\gamma \geq 2$. Since this ratio is always less than 2, we get the limit 2.

On the other hand, there are plenty of examples for which logarithmic efficiency is verified and not BRE, just by having the same exponential decreasing rate for the second moment and square expectation, but with an additional (polynomial) multiplicative component for the second moment, vanishing for logarithmic efficiency, but not for relative error. Other more practical examples, from queuing analysis and large-deviations theory, can be found in [12]. A simpler basic example is provided in [7], just by looking at an estimator for which $\gamma = e^{-\eta/\varepsilon}$ with $\eta > 0$, but for which the variance is $Q(1/\varepsilon)e^{-2\eta/\varepsilon}$ with Q a polynomial.

Extensions of logarithmic efficiency and BRE were introduced in [7] to higher moments than just the second, to make sure that they are well estimated too. For example, this also allows the variance of the empirical variance to be controlled. A preliminary work on this was [17], where BRE for the empirical variance was studied. In Section 4.4, we further investigate the asymptotic coverage of the confidence interval as $\varepsilon \to 0$.

4.3 Efficiency (or work-normalized variance) analysis

Throughout the above analysis, we have been looking at estimators for which the (relative) variance is as small as possible for a fixed sample size. On the other hand, this improved precision might be attained at the cost of employing a more complex algorithm, which can lead to increased computation time. This variation might also depend on the rarity parameter ε. Similarly, some methods can have an average computation cost decreasing with ε. This trade-off between accuracy and computational complexity has therefore to be taken into account with when analyzing rare event simulators.

The principle is then to combine variance and computation time. In [5], the efficiency is defined as being inversely proportional to the product of the sampling variance and the amount of labor required to obtain this estimate. Formally:

Definition 3. *The efficiency of an estimator $\widehat{\gamma}_n$ based on a sample of size n, with variance σ_n^2 and obtained, on average, in a computation time t_n, is $1/(\sigma_n^2 t_n)$.*

If the estimate is obtained from n independent replications each of variance σ^2 and with sampling average time t, then $\sigma_n^2 = \sigma^2/n$ and $t_n/n \to t$ as $n \to \infty$. Thus, if $n \gg 1$, the efficiency of $\widehat{\gamma}$ is approximately $1/(\sigma^2 t)$. This means that $\sigma_n^2 t_n$

can be also seen as a work-normalized variance. It also allows two estimators to be compared for a given computation budget c: if t and t' are the mean times required to generate one independent replication of X and X' when computing $\widehat{\gamma}_n$ and $\widehat{\gamma}'_{n'}$, the number of replications will be respectively $n = c/t$ and $n' = c/t'$. Thus the best estimator is $\widehat{\gamma}_n$ if $\sigma^2(X)t < \sigma^2(X')t'$, that is, if its efficiency is larger.

This definition is generalized in [4] by looking more precisely at the variance obtained with a budget c, taking into account the random generation time.

Based on this principle, the so-called bounded relative efficiency has been defined in [2]:

Definition 4. *Let $\widehat{\gamma}_n$ be an estimator of γ built using n replications and σ_n^2 its variance. Let t_n be the average simulation time to get those n replications. The relative efficiency of $\widehat{\gamma}_n$ is given by*

$$\text{REff} = \frac{\gamma^2}{\sigma_n^2 t_n}.$$

We will say that $\widehat{\gamma}_n$ has bounded relative efficiency with respect to rarity parameter ε, if there exists a constant $d > 0$ such that REff *is minored by d for all ε.*

This basically means that the normalized relative variance $\sigma_n^2 t_n/\gamma^2$ is upper-bounded whatever the rarity, and is therefore a work-normalized version of the bounded relative error property.

In [2], an illustration of the need for such a definition is provided for the reliability analysis of a network (see Chapter 7 below), where the relative error is unbounded but the method is still efficient as $\varepsilon \to 0$, just due to the fact that the *average* computation time per run decreases to 0 at a proper rate. Sufficient conditions for this are also provided.

Similarly, the work-normalized logarithmic efficiency was defined in [3] to deal with the efficiency of splitting estimators.

Definition 5. *The unbiased estimator $\widehat{\gamma}_n$ of γ has work-normalized logarithmic efficiency if*

$$\lim_{\varepsilon \to 0} \frac{\ln t_n + \ln E[\widehat{\gamma}_n^2]}{\ln \gamma} = 2.$$

Note nonetheless that those definitions of relative efficiency and work-normalized logarithmic efficiency are good for comparing the relative merits of two estimators, but are far from perfect definitions. Indeed, there are some flaws in the above definitions. Computing times are usually random, so looking at a fixed computing budget c might be misleading: the number of replications is roughly c/t, but we would need this number to be *uniformly bounded* to make sure that we can bound the error whatever ε. At least, it would be of interest to consider the second moment of the computation time in the definition. This would lead to what could be the valid definition of work-normalized relative error, that is, the relative error for a computing budget c is bounded as $\varepsilon \to 0$. The above definitions, even if informative, are unfortunately more restrictive.

4.4 Another key issue: confidence interval coverage/reliability

Hitherto we have been dealing with the relative error uniformly in ε (or its weaker work-normalized version), but always based on the idea that the coverage of the confidence interval produced by the central limit theorem is always valid. Making sure that the coverage of the confidence interval is uniformly bounded in ε is of interest too.

Similarly, we have highlighted that, because it is the estimated (rather than the exact) variance that is actually used in the confidence interval computation, we may end up with the simple case of an interval $(0, 0)$ because no occurrence of the rare event is detected, but in any case, as illustrated by Section 4.4.1, with an interval for which relative error seems bounded while it is not, and which does not include the exact value. This unpleasant observation highlights the need to design diagnostic procedures in order to point out if we are in this situation and is the focus of Section 4.5. But first, Section 4.4.2 looks at a property asserting the confidence interval coverage validity, while Section 4.4.3 reviews the coverage function representing the actual coverage in terms of the nominal.

4.4.1 Reliability issue of the observed confidence interval

Consider again the illustrative Example 1, with γ estimated by means of $\widehat{\gamma}_n^{\mathrm{IS}}$, where we fix the number n of samples, $n = 10^4$, using the same pseudo-random number generator, and varying ε from 10^{-2} down to 0. Table 4.1 gives, for different values of ε, $2\varepsilon^2$ (the equivalent of γ), $\widehat{\gamma}_n^{\mathrm{IS}}$, the IS estimator, and the 95% confidence interval obtained, together with the *estimated* variance $\widehat{\sigma}_n$. The estimated value becomes bad as $\varepsilon \to 0$: observe that $\widehat{\gamma}_n^{\mathrm{IS}}$ seems to be close to the expected value for $\varepsilon \geq 2 \times 10^{-4}$, and that the confidence interval seems suitable too, but, between 2×10^{-4} and 1×10^{-4}, as ε decays, the results are far from expectations and $2\varepsilon^2$ is not included in the confidence interval anymore. Actually, in this estimation, some paths important for the estimation of γ and of

Table 4.1 Equivalent $2\varepsilon^2$ of γ, IS estimation $\widehat{\gamma}_n^{\mathrm{IS}}$ of γ, confidence interval and estimated relative error for Example 1 using the failure biasing scheme with $q = 0.8$, for a fixed sample size $n = 10^4$ and different values of ε

ε	$2\varepsilon^2$	$\widehat{\gamma}_n^{\mathrm{IS}}$	Confidence Interval	Est. RE
1e-02	2e-04	2.03e-04	(1.811e-04, 2.249e-04)	1.08e-01
1e-03	2e-06	2.37e-06	(1.561e-06, 3.186e-06)	3.42e-01
2e-04	8e-08	6.48e-08	(1.579e-08, 1.138e-07)	7.56e-01
1e-04	2e-08	9.95e-09	(9.801e-09, 1.010e-08)	1.48e-02
1e-06	2e-12	9.95e-13	(9.798e-13, 1.009e-12)	1.48e-02
1e-08	2e-16	9.95e-17	(9.798e-17, 1.009e-16)	1.48e-02

$\mathrm{Var}(\widehat{\gamma}_n^{\mathrm{IS}})$ (paths whose probability is $\Theta(\varepsilon^2)$ under the original measure) are still rare under the IS measure, leading to wrong estimations.

Let us look at this in some detail. Assume that n is fixed and $\varepsilon \to 0$. At some point, ε will be so small that transitions in $\Theta(\varepsilon)$ (see Figure 4.3) are not sampled anymore (probabilistically speaking). Everything happens as if we were working on the model depicted in Figure 4.4. Let us denote by $\mathcal{P}'_F$ the subset of $\mathcal{P}_F$ whose paths belong to this last chain. The expectation of our estimator will now be, on average,

$$\widehat{\gamma}'_n = \sum_{\pi \in \mathcal{P}'_F} p(\pi) \approx \varepsilon^2$$

and, concerning the variance, we will get, also on average,

$$\frac{1}{n}\left[\sum_{\pi \in \mathcal{P}'_F} \frac{p^2(\pi)}{\widetilde{p}(\pi)} - (\widehat{\gamma}'_n)^2\right] \approx \frac{1-q^2}{nq^2}\varepsilon^4.$$

This leads to a (mean) *observed* RE given by

$$\mathrm{RE} \approx \frac{1.96\sqrt{1-q^2}}{q\sqrt{n}},$$

which is independent of ε. The reader can check that these formulas are coherent with the numerical values observed for $\varepsilon \leq 10^{-4}$. So, this is a case where we know that the relative error of the IS technique used is not bounded when rarity increases, but where we numerically observe exactly the contrary. These problems are much harder to detect than the $(0, 0)$ interval case.

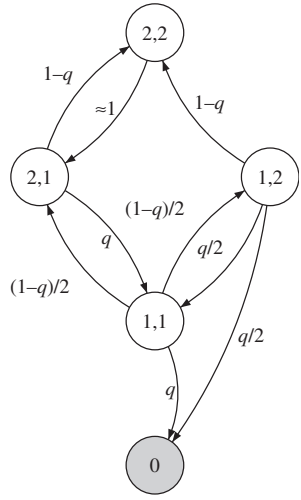

Figure 4.4 Model effectively 'seen' by the IS simulator when transitions in $\Theta(\varepsilon)$ are not observed during n trajectories of the chain.

The question therefore is: what is the validity of the proposed confidence interval? The techniques presented in previous chapters (IS and splitting) consist of different ways to speed up the rare event occurrence, but dealing with the confidence interval coverage might still be an issue.

Consider now the classical $M/M/1/B$ model, where we wish to evaluate $\gamma = \mathbb{P}(\text{reaching } B \text{ before } 0 \mid N(0) = 1)$ (this is Example 3 in Chapter 2), $N(t)$ being the number of customers at time t. More formally:

Example 3. Consider the discrete-time absorbing Markov chain X given in Figure 4.5 and define $\gamma = \mathbb{P}(X(\infty) = B \mid X(0) = 1)$. Observe that this is equal to $\mathbb{P}(\text{reaching } B \text{ before } 0 \mid N(0) = 1)$ in the $M/M/1/B$ queue with arrival rate λ and service rate μ, if $p = \lambda/(\lambda + \mu)$.

This is an elementary example in probability theory, and we know the answer: $\gamma = (r^{-1} - 1)/(r^{-B} - 1)$ if $r = \mu/\lambda = (1 - p)/p \neq 1$ (if $\lambda = \mu$, that is, if $p = 1/2$, then $\gamma = 1/B$). Suppose that we want to estimate γ using the standard simulator. In this example, rarity comes from the combination of values of the parameters p and B, the latter controlling the size of the model, a different situation than in previous example. A typical line of analysis here involves fixing p, varying B, and controlling rarity through $\varepsilon = 1/B$.

For instance, suppose that $p = 0.4$ and $B = 40$. The probability p is not very small, but combined with the size of the chain, we get $\gamma \approx 4.5 \times 10^{-8}$. Suppose we try an IS scheme by simply changing the probability p into some $\widetilde{p} > 1/2$, for instance, $\widetilde{p} = 0.9$, and that we simulate $n = 10^5$ paths of the chain. A standard implementation of this gave the approximate estimate 6.5×10^{-10} and estimated RE $\approx 40\%$. Without knowing the exact value, it is difficult to detect that there is a problem. If we refer to the previous ideas, we can imagine the user increasing B (i.e., increasing rarity), and looking at the behavior of the relative error. In Table 4.2 we provide some numerical results obtained by keeping everything fixed except B, which we increase.

The user may think that the RE looks bounded (while being pretty large), but observe that the exact value is never included in the observed confidence interval. We can suspect the same problem as before, even if the numerical behavior is not exactly the same. Looking again at the case of $B = 40$, it seems reasonable to try increasing the sample size. Keeping everything fixed except the sample size $n = 10^6$, we get an estimate of 1.62×10^{-9} with a relative error $\approx 39\%$. Again, we can suspect the same phenomenon as for the previous example.

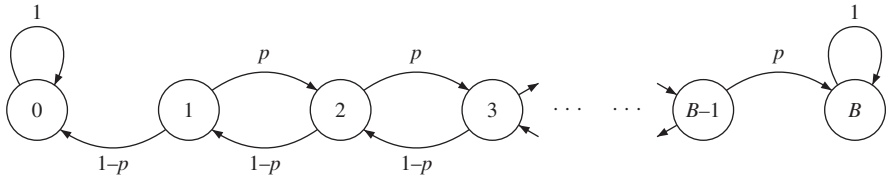

Figure 4.5 Discrete-time Markov chain X associated with the $M/M/1/B$ model, used to compute $\gamma = \mathbb{P}(X(\infty) = B \mid X(0) = 1)$.

Table 4.2 Estimating γ in the $M/M/1/B$ model, with $p = 0.4$, using $n = 10^5$ samples and the failure biasing change of measure with $\widetilde{p} = 0.9$, for different values of the buffer size B. The table gives the exact value of γ, its IS estimate and the estimated RE

B	γ	$\widehat{\gamma}^{\text{IS}}$	Est. RE
40	4.52e-08	6.50e-10	40%
50	7.84e-10	2.46e-12	80%
60	1.36e-11	2.34e-14	120%
70	2.36e-13	1.11e-17	45%
100	1.23e-18	2.21e-24	102%

We observe that in the family of IS methods where the new measure is state-independent (see Chapter 2), the best change of measure for this queue is known: it involves swapping the arrival and the service rate, or equivalently, using $\widetilde{p} = 1 - p$ in discrete time [9]. If we do so, we can check that things go smoothly, and that the estimators behave correctly (no anomaly in the behavior of the RE, nor on the observed likelihood ratio).

The aim of the rest this chapter is to discuss the following questions. How can we define a good estimator? Can it be good whatever the rarity? Can we detect in practice whether an estimate is good or not?

4.4.2 Normal approximation

In [15, 16], the bounded normal approximation (BNA) property is defined, asserting that the Gaussian approximation on which the confidence interval, and thus the confidence interval coverage, is based remains uniformly bounded as ε tends to 0. It finds its roots in the Berry–Esseen theorem which states that if ϱ is the third absolute moment of each of the n independently and identically distributed copies X_i of random variable X (with σ^2 its variance), Φ the standard normal distribution, $\widehat{\gamma}_n = n^{-1}\sum_{i=1}^{n} X_i$, $\widehat{\sigma}_n^2 = n^{-1}\sum_{i=1}^{n}(X_i - \widehat{\gamma}_n)^2$ and F_n the distribution of the centered and normalized sum $(\widehat{\gamma}_n - \gamma)/\widehat{\sigma}_n$, then there exists an absolute constant $a > 0$ such that, for each x and n,

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

Definition 6. *We say that $\widehat{\gamma}_n$ satisfies the bounded normal approximation property if ϱ/σ^3 remains bounded as $\varepsilon \to 0$.*

When this property is satisfied, only a fixed number of iterations are required to obtain a confidence interval having a fixed error no matter the level of rarity.

We could also look at a stricter condition, by making sure that the variance satisfies BRE. This is a stricter condition than BNA because it means looking at

the fourth moment divided by the square of the variance, and, from the Jensen inequality, BRE for the variance implies BNA [17].

In [15], an example is given where BRE is satisfied, but not BNA, so the coverage of the confidence interval is not validated. BRE is therefore not sufficient alone to guarantee the robustness of a rare event estimator.

Note that BNA is a *sufficient* condition for coverage certification, and not a necessary one [15]. For instance, there exist more general versions of the Berry–Esseen bound (see [10]) for which the moment of order $2 + \delta$ is used (with $\delta > 0$) instead of the third moment, being then less restrictive. Note nonetheless that this is at the expense of the convergence rate to the Gaussian distribution, $O(n^{-\delta/2})$ instead of $O(n^{-1/2})$. A generalized version of BNA property could then be as follows:

Definition 7. *We say that $\widehat{\gamma}_n$ satisfies bounded normal approximation if there exists $\delta > 0$ such that $E[|X - \gamma|^{2+\delta}]/\sigma^{2+\delta}$ remains bounded as $\varepsilon \to 0$.*

4.4.3 Coverage function

In order to more directly investigate the actual coverage of confidence intervals for small values of ε when the number of replications is fixed, we can look at the so-called *coverage function* defined by L.W. Schruben in [13]. Define

$$R(\eta, \mathbb{X}) = \left(\widehat{\gamma}_n - c_\eta \frac{\widehat{\sigma}_n}{\sqrt{n}}, \widehat{\gamma}_n + c_\eta \frac{\widehat{\sigma}_n}{\sqrt{n}} \right)$$

as the confidence interval at confidence level η obtained using data

$$\mathbb{X} = (X_i)_{1 \le i \le n}$$

(i.e., $c_\eta = \Phi^{-1}((1 + \eta)/2)$). Under normality assumptions, it is easy to show that $\mathbb{P}[\gamma \in R(\eta, \mathbb{X})] = \eta$. Now define the random variable

$$\eta^* = \inf\{\eta \in [0, 1] : \gamma \in R(\eta, \mathbb{X})\}.$$

η^* should be uniformly distributed, that is,

$$F_{\eta^*}(\eta) = \mathbb{P}[\eta^* \le \eta] = \eta.$$

Not satisfying normal assumptions leads to two potential sources of error:

- $F_{\eta^*}(\eta) < \eta$ may lead to wrong conclusions (lower coverage),
- while if $F_{\eta^*}(\eta) > \eta$ the method is not efficient because a smaller sample size could have been used to get the desired coverage.

In order to investigate the actual coverage function, one can consider independent blocks of data $\mathbb{X} = (X_i)_{1 \le i \le n}$, producing independent realizations of η^*, from which its empirical distribution can be deduced. Reproducing it for different values of ε and looking at deviations from the uniform distribution illustrates the robustness of the estimator. This will be helpful below when discussing possible diagnostic-oriented approaches.

4.5 Diagnostics ideas

This section discusses the issue of detecting potential problems associated with the reliability of rare event confidence intervals. We will review three ideas to deal with these problems from a diagnostic point of view. First, we will see that using the fact that the expectation of the likelihood ratio equals unity in an importance sampling situation, a relevant idea a priori, is actually not of value when dealing with rare events. Second, we look at the possible numerical anomalies that can occur when looking at the behavior of the relative error as the system becomes rarer. A last diagnostic possibility is to make use of the covering function, that is, to look at how far the empirical coverage function is from the uniform.

4.5.1 Checking the value of the expected likelihood ratio

How should a test concerning the reliability of the confidence interval be constructed? A first thought would be to look at properties of the likelihood ratio when dealing with IS. Consider the expected value of a random variable X under probability measure $\mathbb{P}$. IS generates an unbiased estimator by using an IS measure $\widetilde{\mathbb{P}}$ with $d\widetilde{\mathbb{P}} \not\equiv 0$ when $X d\mathbb{P} \not\equiv 0$. Indeed, we then have $\widetilde{\mathbb{E}}[XL] = \mathbb{E}[X] = \gamma$ with $L = d\mathbb{P}/d\widetilde{\mathbb{P}}$ the likelihood ratio (see Chapter 2). We can then easily see that, with the more stringent condition that $d\widetilde{\mathbb{P}} \not\equiv 0$ when $d\mathbb{P} \not\equiv 0$, the expected value of the likelihood ratio is exactly 1. We will assume that this condition is satisfied for the remainder of this subsection, but remark that it is not true in general since we can construct unbiased IS estimates of γ for which $d\widetilde{\mathbb{P}} = 0$ when $X = 0$, such as the zero-variance change of measure.

This observation on the expected value of L could be thought to be a basis for designing a diagnostic: at the same time as we perform the computations needed to construct $\widehat{\gamma}_n^{\mathrm{IS}}$ and the associated confidence interval, we do the same for estimating $\widetilde{\mathbb{E}}[L]$. If the confidence interval obtained does not contain the exact value 1 under the condition that $d\widetilde{\mathbb{P}} \not\equiv 0$ when $d\mathbb{P} \not\equiv 0$, one has to exercise caution.

Why does this diagnostic not work in general? Let X be the the indicator function of a rare set A, that is, $\gamma = \mathbb{E}[\mathbb{1}(A)] = \widetilde{\mathbb{E}}[L\mathbb{1}(A)]$. Then, defining A^c as the complementary set of A and from the expected value of the likelihood ratio, we get

$$1 = \widetilde{\mathbb{E}}[L\mathbb{1}(A)] + \widetilde{\mathbb{E}}[L\mathbb{1}(A^c)] = \gamma + \widetilde{\mathbb{E}}[L\mathbb{1}(A^c)].$$

In order to use a test based on $\widetilde{\mathbb{E}}[L] = 1$, the variance of L has to be small enough so that we do not encounter the aforementioned problems where its variance is underestimated because the second moment has large values with small probability (so that those cases are not reached for a small to moderate sample size n) under $\widetilde{\mathbb{P}}$, and small vales with high probability. Therefore, $\widetilde{\text{Var}}[L\mathbb{1}(A^c)]/n$ has to be small. This is unfortunately not the case in general because the IS scheme is designed to have a small variance for random variable $L\mathbb{1}(A)$, not for L. We indeed have $L \ll 1$, very small on A to be as close as possible to the value of γ for reducing the variance of the estimator, but $L \gg 1$ is likely to happen at some values in A^c.

The next example illustrates this problem of a properly designed IS scheme for which such a test is not going to work well.

Example 4. Consider a random walk $S_n = X_1 + \cdots + X_n$ on the integers or on the reals, starting from 0, where the X_i are independently and identically distributed with cumulative distribution function F. We wish to estimate the probability γ of reaching a level $b > 0$ before a level $-k < 0$. It is assumed that the random walk has a negative drift, meaning that the probability of going up, $X_i > 0$, is smaller that that of going down, $X_i < 0$, leading to a small value of γ. A class of IS measures, called *exponential twisting*, makes use of large deviations (see Chapter 5 for more details on the application of large-deviations theory to random walks). The exponentially twisted IS measure involves replacing dF by

$$d\widetilde{F} = \frac{e^{\theta x}}{M(\theta)} dF(x)$$

with $M(\theta) = \mathbb{E}[e^{\theta X_1}]$, the moment generation function of the X_i. It is known that there exists a θ^* for which $M(\theta^*) = 1$, and that this IS scheme yields logarithmic efficiency. Let us now investigate more closely the behavior of the likelihood ratio. On the paths for which b is reached before $-k$, we have $L \approx e^{-\theta^* b}$, while $L \approx e^{\theta^* k}$ on paths for which $-k$ is reached before, with probability of the order of $e^{-\theta^* k}$.

Now, if the sample size $n \ll e^{\theta^* k}$, we will therefore end up with an estimation of $\widetilde{\mathbb{E}}[L] \approx e^{-\theta^* b}$ because $-k$ is unlikely to be reached, with a small sample variance too. Worse, to get an estimate around 1 as expected, we need $n \gg e^{\theta^* k}$, which can take a longer time if $k > b$ than in the case of crude Monte Carlo, for which n has to be larger than $e^{\theta^* b}$ on the average.

Another interesting remark arises from looking at Example 1. In that case, estimating the expected value of the likelihood ratio always provides a confidence interval for this expectation that includes 1. For instance, using the same numerical values as in Table 4.1, varying ε in the same way, we get for the mean likelihood ratio under the IS measure almost the same confidence interval (0.99, 1.07). A test is therefore not able to detect the difficulty of estimating γ for that kind of example. It is actually the opposite problem than in previous example: it does not provide a warning even if it should, while for the random walk example, it provides an irrelevant warning.

4.5.2 Observed relative error behavior

In Section 4.4.1 we discussed the fact that, in some cases, the simulation technique can degrade when rarity increases, but the numerical values coming from the simulation run hides this phenomenon, leading the user to accept incorrect results. We illustrated this by means of an example where rarity is parameterized by ε, and where in spite of the fact that RE is unbounded as $\varepsilon \to 0$, we will *necessarily* observe that RE suddenly becomes essentially constant, that is, independent of ε. Of course, this is not a systematic fact appearing in these contexts, but it simply underlines the necessity of being careful if we observe this type of behavior.

Specifically, as a diagnostic rule, the idea is to simulate (with small sample sizes) the network for different values of ε larger than in the original problem, that is, to simulate much less rare events, with a small and fixed sample size, before running the 'real' simulation if things seem to go well. What is the incorrect behavior we try to detect? We look to see whether the estimated relative variance seems first to increase, then suddenly drops and stays fixed. This is due to the fact that important events (or paths, depending on the context) in terms of contribution to the variance (and to the estimation itself), are not sampled anymore. This trend of regular growth and sudden drop is likely to be a good hint of rare event problems.

An illustration of this was provided by Example 1, Table 4.1. If we use a sample size $n = 1000$, ten times smaller than that used in Table 4.1, we observe the same phenomenon, always coherent with the formulas given in Section 4.4.1. We observed the same behavior with different configurations.

This type of phenomenon does not appear in the case of the $M/M/1/B$ model presented in Example 3. Increasing B (see Table 4.2), we observe fluctuations of the relative error, but no trend similar to that exhibited before. The diagnostic can hardly be conclusive in this model, as it was in the first example. This illustrates that the tests of the section are traditional *rejection* tests.

Let us now consider another example [1]:

Example 5. Consider the discrete-time Markov chain X given in Figure 4.6 and define $\gamma = \mathbb{E}(K^{-1} \sum_{k=1}^{K} 1(X(k) = 1) \mid X(0) = 1)$.

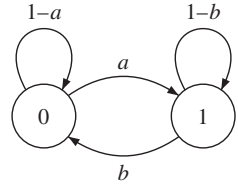

Figure 4.6 A two-state Markov chain. We look at the average fraction of interval $\{1, 2, \ldots, K\}$ *where the chain is in state 1, starting at state 0 at time 0.*

The exact value of γ is

$$\gamma = \frac{a}{a+b}\left[1 - (1-a-b)\frac{1-(1-a-b)^K}{K(a+b)}\right].$$

Assume nevertheless that we use IS to estimate γ and let us consider the cases where

$$P = \begin{pmatrix} 0.99 & 0.01 \\ 0.1 & 0.9 \end{pmatrix}, \quad \widetilde{P} = \begin{pmatrix} 0.4 & 0.6 \\ 0.5 & 0.5 \end{pmatrix}.$$

We look for the value of γ when $K = 30$. We know that the exact answer is $\gamma \approx 6.713^{-2}$, and, of course, this is easy to estimate with the crude estimator. We used the proposed IS scheme for $n = 10^5$ samples, changing the seed of the pseudo-random number generator. We got the results shown in Table 4.3. We can observe here that over these six runs, the relative error fluctuates without a clear trend, but in five of the six cases, the exact value is outside the confidence interval (the case where the exact value is in the confidence interval is for seed).

If we increase the value of K, increasing the possible number of paths, we get the results given in Table 4.4. The RE exhibits no trend again, but we know that the estimations are horribly bad, and that the exact value is never inside the obtained confidence intervals.

In conclusion, for this test, involving checking the behavior of the relative error as a function of rarity, for small sample sizes, we observe good results when rarity is associated with transitions and the state space has a fixed topology, and no clear indications when rarity comes from the increasing length of good paths, as in the $M/M/1/B$ case.

Table 4.3 Estimating γ (whose value is 6.713^{-2}) in the two-state Markov chain of Example 5, with $a = 0.01$, $b = 0.1$, $\widetilde{a} = 0.6$, $\widetilde{b} = 0.5$, $K = 30$, for different seeds (using drand48( ) under Unix), for $n = 10^5$ samples

seed	314159	31415	3141	314	31	3
$\widehat{\gamma}_n^{\mathrm{IS}}$	1.949e-04	1.583e-04	1.282e-04	2.405e-01	6.089e-05	1.021e-04
RE	9.636e-01	8.637e-01	1.667e00	1.958e00	1.263e00	9.686e-01

Table 4.4 Estimating γ in the two-state Markov chain of Example 5, with $a = 0.01$, $b = 0.1$, $\widetilde{a} = 0.6$, $\widetilde{b} = 0.5$, for different values of K, using $n = 10^5$ samples and the same seed (272, with drand48( ) under Unix)

K	30	50	70	90
γ	6.713e-02	7.624e-02	8.040e-02	8.274e-02
$\widehat{\gamma}_n^{\mathrm{IS}}$	4.099e-05	1.748e-10	4.104e-14	2.554e-23
RE	8.723e-01	1.189e00	1.937e00	1.433e00

4.5.3 Diagnostic based on the coverage function

A last diagnostic possibility is to make use of Schruben's coverage function. The algorithm can be described as follows, as hinted in the description of the coverage function. Befored starting to run the (real) simulation, consider smaller sample sizes n and k values of ε, the rarity parameter, $\{\varepsilon_j; 1 \leq j \leq k\}$ with $\varepsilon_1 > \cdots > \varepsilon_k$. For each value ε_j, m independent blocks of data $\mathbb{X} = (X_i(\varepsilon))_{1 \leq i \leq n}$ are then used, giving independent realizations of η^*. From those m realizations, the empirical distribution of η^* can be obtained and compared with the uniform distribution. Then one can see if there is a trend: if the empirical distribution gets farther from the uniform as ε_j decreases, the current estimator can be considered as non-robust (unreliable), and a better one should be chosen. Otherwise, the estimator is not rejected by the test.

An important remark is that, in order to apply this diagnostic, the exact value (or at least an equivalent as $\varepsilon \to 0$) has to be known for computing η^*. As a consequence, the diagnostic can only be used for small instances of the problem. For example, when estimating the probability in an $M/M/1$ queue that the occupancy exceeds a value B (with B large), the exact value can be estimated for smaller values of B, and a trend can be derived. The same applies when dealing with a Markov chain on a small state space, but looking at long simulation times T (such as in Example 5 above), by looking at smaller values of T. The case of large Markov chains where rarity comes from rare transitions is more difficult. But one can try to construct a smaller instance of the model, with similar topology or properties (we do not care about the result being the same) and for which the exact value is known, and look to see whether the coverage function does not deviate as critical transition probabilities decrease. Our three examples describe those three situations and are detailed now.

Figure 4.7 displays the coverage function for the $M/M/1$ queue, looking at the probability that B is reached before returning to 0. This is done for sample sizes $n = 1000$ and repeated $k = 500$ times in order to get the empricial distribution function (smoothed thanks to interpolation). In the numerical experiments, $p = 0.3$ and we chose $\widetilde{p} = 0.5$ (not the optimal value, but to illustrate the behavior). It can be seen that as B increases, the coverage function gets worse and worse, so the estimator is not good here.

Look now at the case of the 2×2 matrix of Example 5, with transition matrices

$$P = \begin{pmatrix} 0.2 & 0.8 \\ 0.2 & 0.8 \end{pmatrix}, \quad \widetilde{P} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}.$$

Again, we take $n = 1000$ and $k = 500$. From Figure 4.8, it can be checked that as the length K of the simulation path increases, the coverage function gets worse and worse, illustrating the bad estimation.

We close our numerical illustrations with Example 1. Figure 4.9 displays the empirical coverage function for different values of ε, still with $n = 1000$ and $k = 500$. Again, as ε decreases, the coverage function gets farther from the uniform, denoting an undesirable behavior.

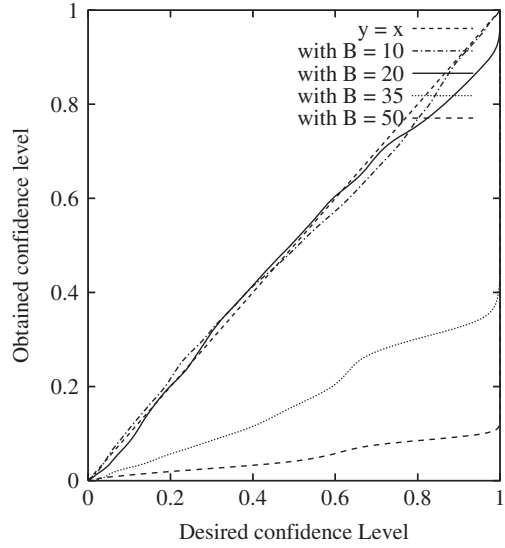

Figure 4.7 Coverage function for the simulation of the $M/M/1$ queue when looking at the probability of exceeding threshold B.

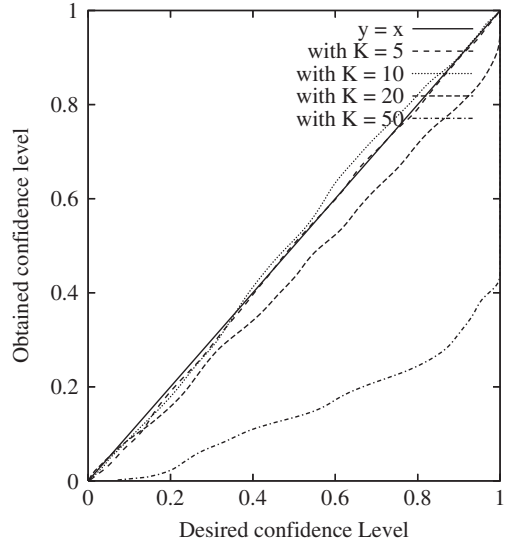

Figure 4.8 Coverage function for the simulation of a two-state Markov chain, as the length K of simulation increases.

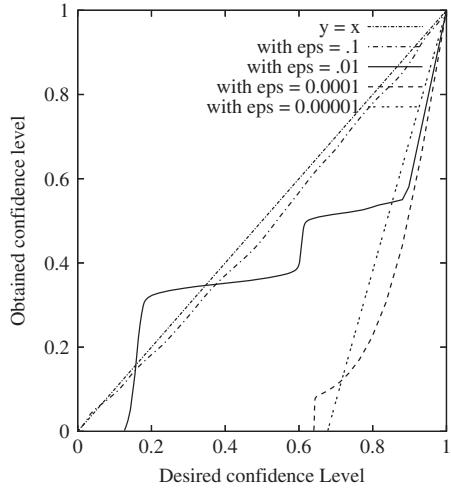

Figure 4.9 Coverage function for Example 1 and various values of ε.

4.6 Conclusions

We discussed the robustness properties (i.e., relative error behavior as the probability of the event goes to zero) we must require an estimator to satisfy when dealing with rare events in a simulation. Together with an overview of these properties and their relations, this chapter also underlined less known problems the practitioner may encounter in this area, concerning the reliability of the confidence interval. The typical situation is a numerical evaluation that can be taken as correctly done, while actually the output of the simulation procedure is completely off target. One of the aims of this chapter is to discuss possible ways of coping with this situation, and to suggest lines of research to derive rules that can be used as diagnostic methods mainly leading to a 'warning' signal along the lines of 'the results of the simulation are suspicious, take care'. But what if such a signal is received? The best advice is to try a different method, or a different parameterization of the technique used.

We concentrated our examples on importance sampling procedures, since this is the most used technique for rare event analysis, and also because it is the one most studied. Observe that the problems underlined here are related to rarity, not just to importance sampling. Also, the rules for detecting problems proposed in this chapter are valid for acceleration methods other than IS-based ones (except obviously for the use of the expected likelihood ratio).

References

[1] S. Andradóttir, D. P. Heyman, and T. J. Ott. On the choice of alternative measures in importance sampling with Markov chains. *Operations Research*, **43**(3): 509–519, 1995.

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

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

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

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

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

[7] P. L'Ecuyer, J. Blanchet, B. Tuffin, and P. W. Glynn. Asymptotic robustness of estimators in rare-event simulation. Technical Report 6281, INRIA, September 2007.

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

[9] S. Parekh and J. Walrand. Quick simulation of rare events in networks. *IEEE Transactions on Automatic Control*, **34**: 54–66, 1989.

[10] V. V. Petrov. *Limit Theorems in Probability Theory*. Oxford University Press, Oxford, 1995.

[11] T. D. Ross. Accurate confidence intervals for binomial proportion and Poisson rate estimation. *Computers in Biology and Medicine*, **33**(3): 509–531, 2003.

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

[13] L. W. Schruben. A coverage function for interval estimators of simulation response. *Management Science*, **26**(1): 18–27, 1980.

[14] P. Shahabuddin. Importance sampling for the simulation of highly reliable Markovian systems. *Management Science*, **40**(3): 333–352, 1994.

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

[16] B. Tuffin. On numerical problems in simulations of highly reliable Markovian systems. In *Proceedings of the 1st International Conference on Quantitative Evaluation of SysTems (QEST)*, pp. 156–164. IEEE Computer Society Press, Los Alamitos, CA, 2004.

[17] B. Tuffin, W. Sandmann, and P. L'Ecuyer. Robustness properties in simulations of highly reliable systems. In *Proceedings of RESIM 2006*, University of Bamberg, Germany, October 2006.