

# ON ASYMPTOTICALLY EFFICIENT SIMULATION OF LARGE DEVIATION PROBABILITIES

A. B. DIEKER,\* \*\* CWI and University of Twente

M. MANDJES,\* \*\*\* CWI and University of Amsterdam

## Abstract

Let  $\{\nu_\varepsilon, \varepsilon > 0\}$  be a family of probabilities for which the decay is governed by a large deviation principle, and consider the simulation of  $\nu_{\varepsilon_0}(A)$  for some fixed measurable set  $A$  and some  $\varepsilon_0 > 0$ . We investigate the circumstances under which an exponentially twisted importance sampling distribution yields an asymptotically efficient estimator. Varadhan's lemma yields necessary and sufficient conditions, and these are shown to improve on certain conditions of Sadowsky. This is illustrated by an example to which Sadowsky's conditions do not apply, yet for which an efficient twist exists.

*Keywords:* Asymptotic efficiency; importance sampling; large deviation

2000 Mathematics Subject Classification: Primary 65C05

Secondary 60F10; 60K10

## 1. Introduction

Given a probability distribution  $\nu$ , we are interested in estimating a rare-event probability  $\nu(A)$ . In direct Monte Carlo methods, the usual estimator is the proportion of times that  $A$  occurs in a certain number of independent samples from  $\nu$ . However, an inherent problem of this approach is that many samples are needed to obtain a reliable estimate for  $\nu(A)$ . In fact, the simulation time required to estimate  $\nu(A)$  may exceed any reasonable limit.

As an important special case, direct Monte Carlo methods are inappropriate for simulating large deviation probabilities. A family of probability measures  $\{\nu_\varepsilon, \varepsilon > 0\}$  is said to satisfy a *large deviation principle* (LDP) if, for a wide class of sets  $A$ ,  $\nu_\varepsilon(A)$  decays exponentially as  $\varepsilon \rightarrow 0$ . Given such a family, we refer to a probability of the form  $\nu_{\varepsilon_0}(A)$ , for some  $\varepsilon_0 > 0$  and some event  $A$ , as a *large deviation probability*. Probabilities of this type are encountered in many fields, e.g. statistics, operations research, information theory, and financial mathematics.

A technique widely used to estimate rare-event probabilities is *importance sampling*. In importance sampling, one samples from a probability measure  $\lambda$ , different from  $\nu_{\varepsilon_0}$ , such that the  $\nu_{\varepsilon_0}$ -rare event becomes  $\lambda$ -likely. Often, one chooses a so-called *exponentially twisted* distribution for  $\lambda$ , but within this class there is still freedom to select a specific twisted distribution. To evaluate the changes of measure, efficiency criteria have been developed. In this paper, we use the *asymptotic efficiency* criterion.

---

Received 7 October 2004; revision received 26 January 2005.

\* Postal address: CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands.

\*\* Email address: ton.dieker@cwi.nl

Supported by the Netherlands Organization for Scientific Research (NWO) under grant 631.000.002.

\*\*\* Email address: michel.mandjes@cwi.nl

Research initiated in the seminal paper [27] has shown that exponentially twisting is asymptotically efficient in specific cases. We mention, in particular, the estimation of the ‘level-crossing’ probability  $P(X_1 + \dots + X_n > M$  for some  $n$ ) for real-valued, independent, and identically distributed (i.i.d.) random variables  $X_1, X_2, \dots$ ; see Lehtonen and Nyrhinen [23], who studied the regime  $M \rightarrow \infty$ . Related results in a more general Markovian setting were obtained by Asmussen [1] and Lehtonen and Nyrhinen [22]. Collamore [8] extended this to a multidimensional setting.

Another example for which an exponential twist is known to yield asymptotic efficiency relates to the ‘Cramér-type’ probability  $P(X_1 + \dots + X_n \geq \gamma n)$  for  $\gamma > E(X_1)$ , where  $n \rightarrow \infty$ . In the case in which  $X_1, X_2, \dots$  have a special Markovian structure, this was found by Bucklew *et al.* [6]. Sadowsky [24] focused on stability issues in the special case of i.i.d. random variables. To estimate  $P(S_n \geq \gamma n)$  for generally distributed  $S_n$ , Sadowsky and Bucklew [26] showed that there exists an asymptotically efficient exponential twist if the Gärtner–Ellis theorem applies to  $\{S_n/n\}$ ; this was also observed by Szechtman and Glynn [28].

However, it has been noted that a successful application of an importance sampling distribution based on large deviation theory critically depends on the specific problem at hand. Glasserman and Wang [18] gave variations on both the level-crossing problem and the Cramér-type problem, and showed that exponential twists can be inefficient if the rare event  $A$  is irregular. In fact, they obtained the stronger result that the so-called relative error can even become unbounded in these examples. Similar observations were made by Glasserman and Kou [16] in a queueing context.

Given the examples of efficient and inefficient simulation with exponentially twisted importance sampling distributions, it is natural to ask whether there exist necessary and sufficient conditions for asymptotic efficiency. In the case in which the Gärtner–Ellis theorem applies, this question was studied by Sadowsky and Bucklew [26], while Sadowsky [25] extended their findings to a general abstract large deviation setting.

The necessary and sufficient conditions presented here have two advantages over those in [25]. The first is that the proof is elementary; the conditions follow straightforwardly from an application of Varadhan’s integral lemma. Therefore, we refer to these conditions as *Varadhan conditions*. Notice that this elementary lemma has previously been applied to derive efficiency properties of certain rare-event estimators (see [14], [17], and [19]). The main result of this paper is that the elementary Varadhan conditions improve on the conditions of [25]. To explain the improvements, it is important to realize that each set of conditions (the Varadhan conditions and Sadowsky’s conditions) applies only under certain assumptions. The assumptions underlying the Varadhan conditions are less restrictive than those underlying Sadowsky’s, meaning that the conditions apply in more situations; notably, convexity of the large deviation rate function is not required. Furthermore, the Varadhan conditions themselves are ‘better’ than Sadowsky’s conditions, i.e. the Varadhan sufficiency condition is implied by Sadowsky’s sufficiency condition, and vice versa for the necessary condition.

The use of the Varadhan conditions is illustrated by an example to which Sadowsky’s results cannot be applied. In this example, asymptotically efficient simulation is possible, while a different approach, which seems perhaps more natural, turns out to be slower.

The paper is organized as follows. After providing the necessary preliminaries in Section 2, we state the Varadhan conditions in Section 3. In Section 4, we show that these conditions improve upon those of Sadowsky. Our example is discussed in Section 5, and in Section 6 we relate the use of a single exponential twist to other approaches.

## 2. Preliminaries

This section provides the basic background on importance sampling and asymptotic efficiency, and discusses their relationship with large deviation techniques. For a more detailed discussion on importance sampling and asymptotic efficiency, see [3], [21], and references therein. Valuable sources for large deviation techniques are the books by Dembo and Zeitouni [9] and Deuschel and Stroock [10].

### 2.1. Importance sampling

Let  $\mathcal{X}$  be a topological space, equipped with some  $\sigma$ -field  $\mathcal{B}$  containing the Borel  $\sigma$ -field. Given a probability measure  $\nu$  on  $(\mathcal{X}, \mathcal{B})$ , we are interested in the simulation of the  $\nu$ -probability of a given event  $A \in \mathcal{B}$ , where  $\nu(A)$  is small. The idea of importance sampling is to sample from a different distribution on  $(\mathcal{X}, \mathcal{B})$ , say  $\lambda$ , in which  $A$  occurs more frequently. This is done by specifying a measurable function  $d\lambda/d\nu: \mathcal{X} \rightarrow [0, \infty]$  and by setting

$$\lambda(B) := \int_B \frac{d\lambda}{d\nu} d\nu.$$

Since  $\lambda$  must be a probability measure,  $d\lambda/d\nu$  should integrate to unity with respect to  $\nu$ .

Assuming equivalence of the measures  $\nu$  and  $\lambda$ , we set  $d\nu/d\lambda := (d\lambda/d\nu)^{-1}$  and note that

$$\nu(A) = \int_A \frac{d\nu}{d\lambda} d\lambda = \int_{\mathcal{X}} \mathbf{1}_A \frac{d\nu}{d\lambda} d\lambda,$$

where  $\mathbf{1}_A$  denotes the indicator function of  $A$ . The importance sampling estimator  $\widehat{\nu_\lambda(A)}$  of  $\nu(A)$  is found by drawing  $N$  independent samples  $X^{(1)}, \dots, X^{(N)}$  from  $\lambda$ ; then

$$\widehat{\nu_\lambda(A)} := \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\{X^{(i)} \in A\}} \frac{d\nu}{d\lambda}(X^{(i)}).$$

It is clear that  $\widehat{\nu_\lambda(A)}$  is an unbiased estimator, i.e.  $E_\lambda(\widehat{\nu_\lambda(A)}) = \nu(A)$ . However, one has the freedom to choose an efficient distribution  $\lambda$ , i.e. a distribution that reduces the variance of the estimator. In particular, it is of interest to find the change of measure that minimizes this variance, or, equivalently,

$$\int_A \left( \frac{d\nu}{d\lambda} \right)^2 d\lambda = \int_{\mathcal{X}} \mathbf{1}_A \left( \frac{d\nu}{d\lambda} \right)^2 d\lambda = \int_A \frac{d\nu}{d\lambda} d\nu.$$

A zero-variance estimator is found by letting  $\lambda$  be the conditional distribution of  $\nu$  given  $A$  (see, e.g. [21]), but it is intractable since  $d\nu/d\lambda$  then depends on the unknown quantity  $\nu(A)$ . This motivates the use of another optimality criterion, namely asymptotic efficiency.

### 2.2. Asymptotic efficiency

To formalize the concept of asymptotic efficiency, we introduce some notions that are extensively used in large deviation theory.

A function  $I: \mathcal{X} \rightarrow [0, \infty]$  is said to be lower semicontinuous if the level sets  $\Phi_I(\alpha) := \{x: I(x) \leq \alpha\}$  are closed subsets of  $\mathcal{X}$  for all  $\alpha \in [0, \infty)$ . The interior and closure of a set  $B \subseteq \mathcal{X}$  are denoted by  $B^\circ$  and  $\bar{B}$ , respectively.

**Definition 1.** A function  $I: \mathcal{X} \rightarrow [0, \infty]$  is called a *rate function* if it is lower semicontinuous. If  $\Phi_I(\alpha)$  is compact for every  $\alpha \geq 0$ ,  $I$  is called a *good rate function*. A set  $B \in \mathcal{B}$  is called an *I-continuity set* if  $\inf_{x \in B^\circ} I(x) = \inf_{x \in B} I(x) = \inf_{x \in \bar{B}} I(x)$ .

The central notion in large deviation theory is the large deviation principle.

**Definition 2.** A family of probability measures  $\{\nu_\varepsilon, \varepsilon > 0\}$  on  $(\mathcal{X}, \mathcal{B})$  satisfies a large deviation principle with rate function  $I$  if, for all  $B \in \mathcal{B}$ ,

$$-\inf_{x \in B^o} I(x) \leq \liminf_{\varepsilon \rightarrow 0} \varepsilon \log \nu_\varepsilon(B) \leq \limsup_{\varepsilon \rightarrow 0} \varepsilon \log \nu_\varepsilon(B) \leq -\inf_{x \in \overline{B}} I(x).$$

Throughout this paper, we assume that the family  $\{\nu_\varepsilon\}$  satisfies an LDP. We fix some *rare event*  $A \in \mathcal{B}$ , i.e.  $\inf_{x \in A^o} I(x) > 0$ , implying that  $\nu_\varepsilon(A)$  decays exponentially as  $\varepsilon \rightarrow 0$ . Since  $\nu(A)$  is supposed to be a large deviation probability, we have  $\nu = \nu_{\varepsilon_0}$  for some  $\varepsilon_0$ .

The definition of asymptotic efficiency is related to the so-called *relative error*. Consider an i.i.d. sample  $X_{\lambda_\varepsilon}^{(1)}, \dots, X_{\lambda_\varepsilon}^{(N)}$  taken from an importance sampling distribution  $\lambda_\varepsilon$ . We define the relative error  $\eta_N(\lambda_\varepsilon, A)$  of the importance sampling estimator

$$\widehat{\nu_{\lambda_\varepsilon}(A)}_N := \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\{X_{\lambda_\varepsilon}^{(i)} \in A\}} \frac{d\nu_\varepsilon}{d\lambda_\varepsilon}(X_{\lambda_\varepsilon}^{(i)})$$

by

$$\eta_N^2(\lambda_\varepsilon, A) := \frac{\text{var}_{\lambda_\varepsilon}(\widehat{\nu_{\lambda_\varepsilon}(A)}_N)}{\nu_\varepsilon^2(A)} = \frac{1}{N} \frac{\text{var}_{\lambda_\varepsilon}(\widehat{\nu_{\lambda_\varepsilon}(A)}_1)}{\nu_\varepsilon^2(A)}.$$

The relative error is proportional to the width of a confidence interval relative to the (expected) estimate itself; hence, it measures the variability of  $\widehat{\nu_{\lambda_\varepsilon}(A)}_N$ .

For asymptotic efficiency, the number of samples required to obtain a prescribed relative error is not allowed to grow exponentially as  $\varepsilon \rightarrow 0$ . Set  $N_{\lambda_\varepsilon}^* := \inf\{N \in \mathbb{N} : \eta_N(\lambda_\varepsilon, A) \leq \eta_{\max}\}$ , for some given maximal relative error  $\eta_{\max}$ , with  $0 < \eta_{\max} < \infty$ .

**Definition 3.** An importance sampling family  $\{\lambda_\varepsilon\}$  is called *asymptotically efficient* if

$$\lim_{\varepsilon \rightarrow 0} \varepsilon \log N_{\lambda_\varepsilon}^* = 0.$$

In the literature, asymptotic efficiency is sometimes referred to as *asymptotic optimality*, *logarithmic efficiency*, or *weak efficiency*. The following lemma provides an equivalent criterion, which is more convenient to work with. Note that it also shows that the specific value of  $\eta_{\max}$  is irrelevant.

**Lemma 1.** An importance sampling family  $\{\lambda_\varepsilon\}$  is asymptotically efficient if and only if

$$\lim_{\varepsilon \rightarrow 0} \varepsilon \log \frac{\int_A (d\nu_\varepsilon / d\lambda_\varepsilon)^2 d\lambda_\varepsilon}{\nu_\varepsilon^2(A)} = 0.$$

*Proof.* First note that, by Jensen’s inequality,

$$\int_A \left( \frac{d\nu_\varepsilon}{d\lambda_\varepsilon} \right)^2 d\lambda_\varepsilon \geq \nu_\varepsilon^2(A),$$

meaning that the criterion in the lemma is equivalent to

$$\limsup_{\varepsilon \rightarrow 0} \varepsilon \log \frac{\int_A (d\nu_\varepsilon / d\lambda_\varepsilon)^2 d\lambda_\varepsilon}{\nu_\varepsilon^2(A)} \leq 0.$$

Using the definition of  $\eta_N(\lambda_\varepsilon, A)$ , we obtain

$$N_{\lambda_\varepsilon}^* = \left\lceil \frac{\text{var}_{\lambda_\varepsilon}(\widehat{v_{\lambda_\varepsilon}(A)})_1}{\eta_{\max}^2 v_\varepsilon^2(A)} \right\rceil.$$

Since  $N_{\lambda_\varepsilon}^* \geq 1$ , we have

$$\limsup_{\varepsilon \rightarrow 0} \varepsilon \log \frac{\int_A (dv_\varepsilon / d\lambda_\varepsilon)^2 d\lambda_\varepsilon}{v_\varepsilon^2(A)} = \limsup_{\varepsilon \rightarrow 0} \varepsilon \log \left( N_{\lambda_\varepsilon}^* + \frac{1}{\eta_{\max}^2} \right) = \limsup_{\varepsilon \rightarrow 0} \varepsilon \log N_{\lambda_\varepsilon}^*,$$

from which the claim can be obtained.

Let us now discuss another characterization of asymptotic optimality, which applies under a weak additional assumption. For this, note that

$$\limsup_{\varepsilon \rightarrow 0} \varepsilon \log \frac{\int_A (dv_\varepsilon / d\lambda_\varepsilon)^2 d\lambda_\varepsilon}{v_\varepsilon^2(A)} \leq \limsup_{\varepsilon \rightarrow 0} \varepsilon \log \int_A \left( \frac{dv_\varepsilon}{d\lambda_\varepsilon} \right)^2 d\lambda_\varepsilon - 2 \liminf_{\varepsilon \rightarrow 0} \varepsilon \log v_\varepsilon(A), \tag{1}$$

with equality if the limit  $\lim_{\varepsilon \rightarrow 0} \varepsilon \log v_\varepsilon(A)$  exists. Sufficient for the existence of this limit is that  $A$  be an  $I$ -continuity set; then  $\lim_{\varepsilon \rightarrow 0} \varepsilon \log v_\varepsilon(A) = -\inf_{x \in A} I(x)$ . In many applications,  $A$  is indeed an  $I$ -continuity set, in which case asymptotic efficiency is equivalent to

$$\limsup_{\varepsilon \rightarrow 0} \varepsilon \log \int_A \left( \frac{dv_\varepsilon}{d\lambda_\varepsilon} \right)^2 d\lambda_\varepsilon \leq -2 \inf_{x \in \bar{A}} I(x).$$

As before, this is seen to be equivalent to  $\lim_{\varepsilon \rightarrow 0} \varepsilon \log \int_A (dv_\varepsilon / d\lambda_\varepsilon)^2 d\lambda_\varepsilon = -2 \inf_{x \in \bar{A}} I(x)$ . The same argument shows that the condition

$$\lim_{\varepsilon \rightarrow 0} \frac{\log \int_A (dv_\varepsilon / d\lambda_\varepsilon)^2 d\lambda_\varepsilon}{\log v_\varepsilon(A)} = 2$$

is equivalent to asymptotic efficiency when  $A$  is an  $I$ -continuity set.

### 3. The Varadhan conditions for efficiency of exponential twisting

In this section, we investigate the asymptotic efficiency of the estimators that are based on an exponential twist. After formalizing the imposed assumptions, we state necessary and sufficient conditions based on Varadhan’s lemma. In Section 4, we discuss their relationship with the conditions developed by Sadowsky [25].

Let  $\mathcal{X}$  be a topological space and  $\mathcal{B}$  be a  $\sigma$ -field on  $\mathcal{X}$  containing the Borel  $\sigma$ -field. We assume that  $\mathcal{X}$  is also a vector space, but not necessarily a topological vector space. Throughout this section, we fix a rare event  $A \in \mathcal{B}$  and a continuous linear functional  $\xi : \mathcal{X} \rightarrow \mathbb{R}$ . Having a topological vector space in mind, we write  $\langle \xi, \cdot \rangle$  for  $\xi(\cdot)$ . We are given a family  $\{v_\varepsilon\}$  of probability measures on  $(\mathcal{X}, \mathcal{B})$ .

**Assumption 1.** (Varadhan assumptions.) *Assume that*

- (i)  $\mathcal{X}$  is a vector space endowed with some regular Hausdorff topology,
- (ii)  $\{v_\varepsilon\}$  satisfies the LDP with a good rate function  $I$ , and

(iii)

$$\lim_{M \rightarrow \infty} \limsup_{\varepsilon \rightarrow 0} \varepsilon \log \int_{\{x \in \mathcal{X} : \langle \xi, x \rangle \geq M\}} \exp\left[\frac{\langle \xi, x \rangle}{\varepsilon}\right] \nu_\varepsilon(dx) = -\infty, \tag{2}$$

and similarly with  $\xi$  replaced by  $-\xi$ .

We note that a simple sufficient condition for (2) to hold is given in Lemma 4.3.8 of [9]:

$$\limsup_{\varepsilon \rightarrow 0} \varepsilon \log \int \exp\left[\frac{\gamma \langle \xi, x \rangle}{\varepsilon}\right] \nu_\varepsilon(dx) < \infty \quad \text{for some } \gamma > 1,$$

and similarly with  $\xi$  replaced by  $-\xi$ .

A new family  $\{\lambda_\varepsilon^\xi\}$  of probability measures is defined by

$$\begin{aligned} \frac{d\lambda_\varepsilon^\xi}{d\nu_\varepsilon}(x) &:= \exp\left[\frac{\langle \xi, x \rangle}{\varepsilon} - \log \int_{\mathcal{X}} \exp\left[\frac{\langle \xi, y \rangle}{\varepsilon}\right] \nu_\varepsilon(dy)\right] \\ &= \frac{\exp[\langle \xi, x \rangle / \varepsilon]}{\int_{\mathcal{X}} \exp[\langle \xi, y \rangle / \varepsilon] \nu_\varepsilon(dy)}. \end{aligned} \tag{3}$$

The measures  $\{\lambda_\varepsilon^\xi\}$  are called *exponentially twisted with twist  $\xi$* . If the family  $\{\lambda_\varepsilon^\xi\}$  is asymptotically efficient, then we simply call the exponential twist  $\xi$  asymptotically efficient.

The following proposition plays a key role in the proofs of this section.

**Proposition 1.** *Let  $d\lambda_\varepsilon^\xi / d\nu_\varepsilon$  be given by (3), and let  $B \in \mathcal{B}$ . Under Assumption 1, we have*

$$\begin{aligned} \liminf_{\varepsilon \rightarrow 0} \varepsilon \log \int_B \left(\frac{d\nu_\varepsilon}{d\lambda_\varepsilon^\xi}\right)^2 d\lambda_\varepsilon^\xi &\geq - \inf_{x \in \mathcal{X}} [I(x) - \langle \xi, x \rangle] - \inf_{x \in B^c} [I(x) + \langle \xi, x \rangle], \\ \limsup_{\varepsilon \rightarrow 0} \varepsilon \log \int_B \left(\frac{d\nu_\varepsilon}{d\lambda_\varepsilon^\xi}\right)^2 d\lambda_\varepsilon^\xi &\leq - \inf_{x \in \mathcal{X}} [I(x) - \langle \xi, x \rangle] - \inf_{x \in B} [I(x) + \langle \xi, x \rangle]. \end{aligned}$$

*Proof.* Fix  $B \in \mathcal{B}$  and note that

$$\begin{aligned} \varepsilon \log \int_B \left(\frac{d\nu_\varepsilon}{d\lambda_\varepsilon^\xi}\right)^2 d\lambda_\varepsilon^\xi &= \varepsilon \log \int_B \frac{d\nu_\varepsilon}{d\lambda_\varepsilon^\xi} d\nu_\varepsilon \\ &= \varepsilon \log \int_{\mathcal{X}} \exp\left[\frac{\langle \xi, x \rangle}{\varepsilon}\right] \nu_\varepsilon(dx) + \varepsilon \log \int_B \exp\left[-\frac{\langle \xi, x \rangle}{\varepsilon}\right] \nu_\varepsilon(dx). \end{aligned} \tag{4}$$

By Assumption 1 and the continuity of the functional  $\xi$ , Varadhan’s integral lemma [9, Theorem 4.3.1] applies. Thus, the limit of the first term exists and equals

$$\lim_{\varepsilon \rightarrow 0} \varepsilon \log \int_{\mathcal{X}} \exp\left[\frac{\langle \xi, x \rangle}{\varepsilon}\right] \nu_\varepsilon(dx) = \sup_{x \in \mathcal{X}} [\langle \xi, x \rangle - I(x)].$$

A similar argument can be applied to the second term in (4). Indeed, we apply a variant of Varadhan’s integral lemma (see, e.g. Exercise 4.3.11 of [9]) to the continuous functional  $-\xi$ : for any open set  $G$  and any closed set  $F$ ,

$$\begin{aligned} \liminf_{\varepsilon \rightarrow 0} \varepsilon \log \int_G \exp\left[-\frac{\langle \xi, x \rangle}{\varepsilon}\right] \nu_\varepsilon(dx) &\geq - \inf_{x \in G} [I(x) + \langle \xi, x \rangle], \\ \limsup_{\varepsilon \rightarrow 0} \varepsilon \log \int_F \exp\left[-\frac{\langle \xi, x \rangle}{\varepsilon}\right] \nu_\varepsilon(dx) &\leq - \inf_{x \in F} [I(x) + \langle \xi, x \rangle]. \end{aligned}$$

In particular, these inequalities hold for  $B^o$  and  $\bar{B}$ . The claim follows by adding the two terms in (4) (using the fact that the limit of the first term exists).

The necessary and sufficient conditions, formulated in the next theorem, follow almost immediately from Proposition 1.

**Theorem 1.** (Varadhan conditions.) *Let Assumption 1 hold. The exponential twist  $\xi$  is asymptotically efficient if*

$$\inf_{x \in \mathcal{X}} [I(x) - \langle \xi, x \rangle] + \inf_{x \in \bar{A}} [I(x) + \langle \xi, x \rangle] \geq 2 \inf_{x \in A^o} I(x). \tag{5}$$

*Let Assumption 1 hold and let  $A$  be an  $I$ -continuity set. If the exponential twist  $\xi$  is asymptotically efficient, then*

$$\inf_{x \in \mathcal{X}} [I(x) - \langle \xi, x \rangle] + \inf_{x \in A^o} [I(x) + \langle \xi, x \rangle] \geq 2 \inf_{x \in \bar{A}} I(x).$$

*Proof.* Sufficiency follows from Lemma 1, the upper bound of Proposition 1, and the LDP of Assumption 1(ii):

$$\begin{aligned} \limsup_{\varepsilon \rightarrow 0} \varepsilon \frac{\int_A (dv_\varepsilon / d\lambda_\varepsilon^\xi)^2 d\lambda_\varepsilon^\xi}{v_\varepsilon^2(A)} &\leq \limsup_{\varepsilon \rightarrow 0} \varepsilon \log \int_A \left( \frac{dv_\varepsilon}{d\lambda_\varepsilon^\xi} \right)^2 d\lambda_\varepsilon^\xi - 2 \liminf_{\varepsilon \rightarrow 0} \varepsilon \log v_\varepsilon(A) \\ &\leq - \inf_{x \in \mathcal{X}} [I(x) - \langle \xi, x \rangle] - \inf_{x \in \bar{A}} [I(x) + \langle \xi, x \rangle] + 2 \inf_{x \in A^o} I(x). \end{aligned}$$

For necessity, the argument is similar. Note that the lower bound of Proposition 1 implies that

$$\limsup_{\varepsilon \rightarrow 0} \varepsilon \log \int_A \left( \frac{dv_\varepsilon}{d\lambda_\varepsilon^\xi} \right)^2 d\lambda_\varepsilon^\xi \geq - \inf_{x \in \mathcal{X}} [I(x) - \langle \xi, x \rangle] - \inf_{x \in A^o} [I(x) + \langle \xi, x \rangle].$$

Furthermore, by the large deviation upper bound,

$$\liminf_{\varepsilon \rightarrow 0} \varepsilon \log v_\varepsilon(A) \leq - \inf_{x \in \bar{A}} I(x).$$

Combining these observations with the assumption that  $A$  is an  $I$ -continuity set, we have equality in (1), so that, by Lemma 1,

$$\begin{aligned} 0 &= \limsup_{\varepsilon \rightarrow 0} \varepsilon \log \int_A \left( \frac{dv_\varepsilon}{d\lambda_\varepsilon^\xi} \right)^2 d\lambda_\varepsilon^\xi - 2 \lim_{\varepsilon \rightarrow 0} \varepsilon \log v_\varepsilon(A) \\ &\geq - \inf_{x \in \mathcal{X}} [I(x) - \langle \xi, x \rangle] - \inf_{x \in A^o} [I(x) + \langle \xi, x \rangle] + 2 \inf_{x \in \bar{A}} I(x), \end{aligned}$$

as desired.

As suggested by the form of Theorem 1, the sufficient condition is also necessary under a weak condition on the set  $A$ . We formalize this in the following corollary, which follows straightforwardly from Theorem 1.

**Corollary 1.** *Let Assumption 1 hold and assume that  $A$  is both an  $I$ -continuity set and an  $(I + \xi)$ -continuity set. Exponentially twisting with  $\xi$  is asymptotically efficient if and only if*

$$\inf_{x \in \mathcal{X}} [I(x) - \langle \xi, x \rangle] + \inf_{x \in \bar{A}} [I(x) + \langle \xi, x \rangle] = 2 \inf_{x \in \bar{A}} I(x).$$

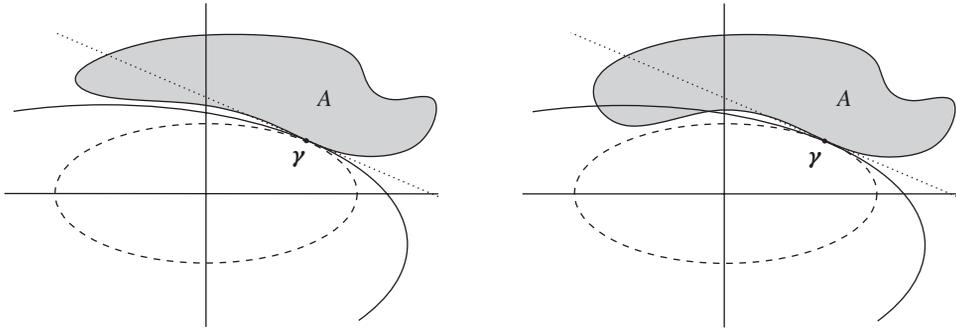


FIGURE 1: Efficient simulation with twist  $\xi_\gamma$  (left) and inefficient simulation with twist  $\xi_\gamma$  (right).

We remark that Sadowsky [25] used a more general notion than asymptotic efficiency, namely  $\nu$ -efficiency. Given an  $I$ -continuity set  $A$ , the importance sampling distribution  $\lambda_\varepsilon^\xi$  is said to be  $\nu$ -efficient if

$$\limsup_{\varepsilon \rightarrow 0} \varepsilon \log \int_A \left( \frac{d\nu_\varepsilon}{d\lambda_\varepsilon^\xi} \right)^\nu d\lambda_\varepsilon \leq -\nu \inf_{x \in A} I(x).$$

In this terminology, we have established conditions for 2-efficiency (see the remarks after Definition 3). To obtain conditions for  $\nu$ -efficiency for general  $\nu \geq 2$ , the statements in this section are easily modified. As an example, when  $A$  is an  $(I + (\nu - 1)\xi)$ -continuity set and when Assumption 1(iii) holds with  $\xi$  replaced by both  $(\nu - 1)\xi$  and  $-(\nu - 1)\xi$ , the exponential twist  $\xi$  is  $\nu$ -efficient if and only if

$$\inf_{x \in X} [I(x) - (\nu - 1)\langle \xi, x \rangle] + \inf_{x \in A} [I(x) + (\nu - 1)\langle \xi, x \rangle] = \nu \inf_{x \in A} I(x).$$

We now illustrate the Varadhan conditions in a simple example. Let  $\nu$  be the distribution of a random variable  $X$  on  $\mathbb{R}^d$ , denote the distribution of the sample mean of  $n$  i.i.d. copies of  $X$  by  $\nu_n$  (note that  $1/n$  plays the role of  $\varepsilon$  in this example), and let  $\nu$  be such that Cramér’s theorem holds.

For instance, take  $\nu_n$  to be a zero-mean bivariate Gaussian distribution with covariance of the form  $\Sigma/n$  for some diagonal matrix  $\Sigma$ ; see Figure 1. We are interested in  $\nu_n(A)$  for two different sets  $A$ ; these are drawn in the left- and right-hand diagrams of Figure 1. Note that the rate function has the form  $I(x_1, x_2) = C_1x_1^2 + C_2x_2^2$  for some constants  $C_1, C_2 > 0$ . As indicated by the dashed level curve of  $I$ , the ‘most likely’ point in  $A$  is in both cases  $\gamma$ , i.e.  $\arg \inf_{x \in A} I(x) = I(\gamma)$ . We can see that there is only one exponential twist  $\xi_\gamma \in \mathbb{R}^2$  of interest for simulation purposes, namely the conjugate point of  $\gamma$ . The level curve of  $I + \xi_\gamma^T \cdot + \inf_{x \in \mathbb{R}^2} [I(x) - \xi_\gamma^T x]$  that goes through  $\gamma$  is depicted as a solid line. Since both sets  $A$  are  $I$ - and  $(I + \xi_\gamma^T \cdot)$ -continuity sets, the twist  $\xi_\gamma$  is asymptotically efficient if and only if  $A$  lies entirely ‘outside’ the solid level curve (see Corollary 1). Hence, in the left-hand diagram the twist  $\xi_\gamma$  is asymptotically efficient, and in the right-hand diagram it is not.

In the literature, sufficient conditions for asymptotic efficiency have been given in terms of dominating points and convexity of  $A$  in the case in which the rate function is convex (see, e.g. [26] and references therein). Using Figure 1, we can explain how it can be seen that the Varadhan conditions improve upon these dominating point conditions (convexity of  $A$  implies the existence of a dominating point, so we focus on dominating points). Every  $I$ -continuity set

that touches  $\gamma$  and is contained in the half-space above the dotted line has dominating point  $\gamma$ . Obviously, such a set lies outside the solid level curve, and one can therefore estimate  $v_n(A)$  asymptotically efficiently by an exponential twist. However, Figure 1 indicates that the dominating point condition is not necessary: neither of the sets  $A$  have a dominating point, while an efficient twist exists in the left-hand diagram.

#### 4. Comparison with Sadowsky’s conditions

General necessary and sufficient conditions for asymptotic efficiency were also developed by Sadowsky [25]. In this subsection, we compare the conditions of Theorem 1 with Sadowsky’s conditions. We show that the assumptions underlying Varadhan’s conditions are less restrictive than Sadowsky’s assumptions. Moreover, the sufficient condition in Theorem 1 improves upon Sadowsky’s sufficiency condition, and the same holds for the accompanying necessary conditions.

In addition to the notation of the preceding subsection, we first introduce some new notions. In this section,  $\mathcal{X}$  denotes a topological vector space, and  $\mathcal{X}^*$  denotes the space of linear continuous functionals  $\xi : \mathcal{X} \rightarrow \mathbb{R}$ . Let  $f : \mathcal{X} \rightarrow (-\infty, \infty]$  be a convex function. A point  $x \in \mathcal{X}$  is called an *exposed point* of  $f$  if there exists a  $\delta \in \mathcal{X}^*$  such that  $f(y) > f(x) + \langle \delta, y - x \rangle$  for all  $y \neq x$ . The functional  $\delta$  is then called an *exposing hyperplane* of  $f$  at  $x$ .

To compare the Varadhan conditions to Sadowsky’s, we first recall Sadowsky’s assumptions.

**Assumption 2.** (Sadowsky’s assumptions.) *Assume that*

- (i)  $\mathcal{X}$  is a locally convex Hausdorff topological vector space,
- (ii)  $\{v_\varepsilon\}$  satisfies the LDP with a convex good rate function  $I$ ,
- (iii) for every  $\delta \in \mathcal{X}^*$ ,

$$\Lambda(\delta) := \limsup_{\varepsilon \rightarrow 0} \varepsilon \log \int_{\mathcal{X}} \exp \left[ \frac{\langle \delta, x \rangle}{\varepsilon} \right] v_\varepsilon(dx) < \infty,$$

and

- (iv)  $A$  satisfies

$$0 < \inf_{x \in A^o \cap \mathcal{F}} I(x) = \inf_{x \in A} I(x) = \inf_{x \in \bar{A}} I(x) < \infty,$$

where  $\mathcal{F}$  denotes the set of exposed points of  $I$ .

Although Assumption 2 looks very similar to Assumption 1, there are crucial differences. To begin with,  $\mathcal{X}$  is not assumed to be a *topological* vector space in Assumption 1(i). To see the importance this difference has for applications, note that the space  $D([0, 1], \mathbb{R})$  of càdlàg functions on  $[0, 1]$  with values in  $\mathbb{R}$  is a (regular Hausdorff) vector space but is not a topological vector space when equipped with the Skorokhod topology. We stress that the regularity of  $\mathcal{X}$  assumed in Assumption 1(i) is implicit in Assumption 2(i): any real Hausdorff topological vector space is regular.

Moreover, we do not assume convexity of the large deviation rate function in Assumption 1(ii). Note that this convexity is given when an LDP is derived using an (abstract) Gärtner–Ellis-type theorem, but that nonconvex rate functions also arise naturally in applications; see Section 5 for a discussion. Assumption 2(iii) implies Assumption 1 since  $\gamma\xi$  is a continuous linear functional for any  $\gamma \in \mathbb{R}$ , while Assumption 2(iv) is slightly stronger than the requirement that  $A$  be an  $I$ -continuity set.

In the above comparison between Assumption 1 and Assumption 2, we have shown the following.

**Proposition 2.** *Assumption 2 implies that Assumption 1 holds and that  $A$  is an  $I$ -continuity set.*

In the remainder of this subsection, we compare the necessary and sufficient conditions of Theorem 1 to the conditions of [25]. Such a comparison is only possible when Sadowsky’s assumptions hold, i.e. we must impose the (stronger) Assumption 2. We start by repeating Sadowsky’s conditions. Given that Assumption 2(iv) holds for  $A$ , we call  $\gamma \in \bar{A}$  a *point of continuity* if  $I(\gamma) = \inf_{x \in \bar{A}} I(x)$  and there exists a sequence  $\{\gamma_n\} \subset A^o \cap \mathcal{F}$  such that  $\gamma_n \rightarrow \gamma$ .

**Theorem 2.** (Sadowsky’s conditions.) *Let Assumption 2 hold. The exponential twist  $\xi$  is asymptotically efficient if*

- (a) *there is a point of continuity  $\gamma$  such that  $I(\gamma) = \langle \xi, \gamma \rangle - \Lambda(\xi)$ ,*
- (b)  *$I(x) + \langle \xi, x \rangle \geq I(\gamma) + \langle \xi, \gamma \rangle$  for all  $x \in \bar{A}$ , and*
- (c) *either  $\langle \xi, x \rangle \geq \langle \xi, \gamma \rangle$  for all  $x \in \bar{A}$ , or there exists an  $x \in \mathcal{F}$  such that  $\xi$  is an exposing hyperplane of  $I$  at  $x$ .*

*Let Assumption 2 hold. If the twist  $\xi$  is asymptotically efficient, then*

- (a) *there is a point of continuity  $\gamma$  such that  $I(\gamma) = \langle \xi, \gamma \rangle - \Lambda(\xi)$ , and*
- (b)  *$I(x) + \langle \xi, x \rangle \geq I(\gamma) + \langle \xi, \gamma \rangle$  for all  $x \in A^o \cap \mathcal{F}$ .*

**Proposition 3.** *Let Assumption 2 hold. The sufficient condition in Theorem 2 implies the sufficient condition in Theorem 1.*

*Proof.* By condition (a) of Theorem 2, there exists a point of continuity  $\gamma \in \bar{A}$  such that  $I(\gamma) = \inf_{x \in \bar{A}} I(x) = \langle \xi, \gamma \rangle - \Lambda(\xi)$ . Since we assume that an LDP holds for some convex  $I$  (Assumption 2(ii)) and that Assumption 2(iii) holds, by Theorem 4.5.10(b) of [9] we have  $I(x) = \sup_{\delta \in \mathcal{X}^*} [\langle \delta, x \rangle - \Lambda(\delta)]$  and, hence,  $I(x) \geq \langle \xi, x \rangle - \Lambda(\xi)$ . Combining this with  $I(\gamma) = \langle \xi, \gamma \rangle - \Lambda(\xi)$ , we conclude that

$$\inf_{x \in \mathcal{X}} [I(x) - \langle \xi, x \rangle] \geq -\Lambda(\xi) = I(\gamma) - \langle \xi, \gamma \rangle,$$

where the inequality may obviously be replaced by an equality.

It is immediate from condition (b) of Theorem 2 that  $\inf_{x \in \bar{A}} [I(x) + \langle \xi, x \rangle] = I(\gamma) + \langle \xi, \gamma \rangle$ . Since  $\inf_{x \in A^o} I(x) = I(\gamma)$ , this implies the sufficient condition (5) in Theorem 1.

It is important to note that we did not use part (c) of Sadowsky’s sufficient condition in the proof of Proposition 3; this part is redundant.

**Proposition 4.** *Let Assumption 2 hold. The necessary condition in Theorem 2 is implied by the necessary condition in Theorem 1.*

*Proof.* Let the twist  $\xi$  be asymptotically efficient. We start by showing that a point of continuity exists under Assumption 2. First note that  $\inf_{x \in A^o \cap \mathcal{F}} I(x) = \inf_{x \in \bar{A}} I(x)$  (Assumption 2(iv)) implies that, for any  $n \in \mathbb{N}$ , we can find some  $\gamma_n \in A^o \cap \mathcal{F} \cap K_n$ , where

$$K_n := \left\{ x \in \mathcal{X} : I(x) \leq \inf_{y \in \bar{A}} I(y) + \frac{1}{n} \right\}.$$

Using the fact that  $\inf_{x \in \bar{A}} I(x) < \infty$  and the rate function is good (Assumption 2(ii)), we see that  $K_n$  is a compact subset of  $\mathcal{X}$  and, hence, is also sequentially compact. Since  $K_n$  decreases in  $n$ , we obviously have  $\{\gamma_n\} \subset K_1$ . Hence, we can substract a subsequence that converges to  $\gamma \in K_1$ , say. Since  $K_n$  is closed for every  $n$  and  $\{\gamma_n\}$  is eventually in  $K_n$ , we must also have  $\gamma \in K_n$  for every  $n$ . As a consequence, we have  $I(\gamma) \leq \inf_{x \in \bar{A}} I(x)$ . Moreover, since  $\{\gamma_n\} \subset A^o \cap \mathcal{F}$ , we also see that  $\gamma \in \overline{A^o \cap \mathcal{F}} \subset \bar{A}$ . Therefore,  $I(\gamma) = \inf_{x \in A^o \cap \mathcal{F}} I(x) = \inf_{x \in \bar{A}} I(x)$  and, so,  $\gamma$  is a point of continuity.

The necessary condition in Theorem 1 implies that

$$\begin{aligned} 2I(\gamma) &\leq \inf_{x \in A^o} [I(x) + \langle \xi, x \rangle] - \sup_{x \in \mathcal{X}} [\langle \xi, x \rangle - I(x)] \\ &\leq \lim_{n \rightarrow \infty} [I(\gamma_n) + \langle \xi, \gamma_n \rangle] - [\langle \xi, \gamma \rangle - I(\gamma)] = 2I(\gamma). \end{aligned}$$

As a result, the inequalities can be replaced by equalities, and we obtain

$$\sup_{x \in \mathcal{X}} [\langle \xi, x \rangle - I(x)] = \langle \xi, \gamma \rangle - I(\gamma) \quad \text{and} \quad \inf_{x \in A^o} [I(x) + \langle \xi, x \rangle] = I(\gamma) + \langle \xi, \gamma \rangle.$$

By Theorem 4.5.10(a) of [9], we also have  $\sup_{x \in \mathcal{X}} [\langle \xi, x \rangle - I(x)] = \Lambda(\xi)$  under Assumption 2. Hence,  $I(\gamma) = \langle \xi, \gamma \rangle - \Lambda(\xi)$  and part (a) of Sadowsky's necessary condition follows. Part (b) is immediate upon noting that having  $\inf_{x \in A^o} [I(x) + \langle \xi, x \rangle] = I(\gamma) + \langle \xi, \gamma \rangle$  implies that  $I(x) + \langle \xi, x \rangle \geq I(\gamma) + \langle \xi, \gamma \rangle$  for all  $x \in A^o$ .

## 5. An example

In this section, we provide an example showing how Corollary 1 is typically used. The conditions of [25] do not apply to this example, since the rate function is nonconvex. Despite this nonconvexity, we show that an exponential twist may still be asymptotically efficient.

Nonconvex rate functions arise naturally in several large deviation settings. Notably, certain large deviations of Markov chains and Markov processes (such as diffusions) are governed by rate functions that need not be convex; see, e.g. [15]. Intuitively, analyzing the rate of convergence of random functions to a nonlinear (deterministic) function causes the rate function to be nonconvex. Nonconvex rate functions also appear when we investigate the rate of convergence to a nondegenerate measure; the example of this section is of the latter type.

In our example, it appears that the event under consideration can be cut into disjoint 'subevents' that comply with Sadowsky's conditions. However, such a 'cutting approach' might be impossible in other cases, or might lead to a large number of subevents (that need to be estimated separately). This is especially relevant to the simulation of hitting probabilities of stochastic processes, as in [8]. Depending on the shape of the hitting curve, simulation with a single twist may work, but the 'cutting approach' leads to the simulation of many events (each of which corresponds to hitting the curve at a particular point in time).

To avoid technicalities that are irrelevant to the paper, we do not illustrate the Varadhan conditions in a sample path setting. Instead, we discuss a relatively simple example, which nevertheless gives a good idea of under what circumstances the Varadhan conditions are more useful than Sadowsky's conditions.

Let us first give some background on our example. Recall that a phase-type distribution is associated with a finite Markov process, and that it can be characterized by three quantities  $(E, \alpha, T)$ ; see, e.g. [2, Chapter III.4]. Given an arbitrary distribution  $\nu$  on  $(0, \infty)$ , we can find a sequence of phase-type distributions that converges weakly to  $\nu$  [2, Theorem III.4.2].

This implies that  $\nu_n(A) \rightarrow 0$  for a large number of sets  $A$ ; in fact,  $\nu_n(A)$  then vanishes at an exponential rate. We are interested in  $\nu_n(A)$  for a fixed  $n$ .

We consider a particularly simple distribution, namely one that is concentrated on  $\{1, 5\}$ . We write  $\alpha = \nu(\{1\}) = 1 - \nu(\{5\})$ . It is notationally cumbersome to describe a sequence of phase-type distributions that converges weakly to  $\nu$  in the  $(E, \alpha, T)$ -notation; a direct description is more appropriate here. Define  $\nu_n$  to be the distribution of  $Y_n$ , where  $Y_n$  has an Erlang( $n, n$ ) distribution with probability  $\alpha$  and an Erlang( $n, \frac{1}{5}n$ ) distribution with probability  $1 - \alpha$ . (Recall that the sum of  $k$  independent exponentially distributed random variables with parameter  $\lambda$  has an Erlang( $k, \lambda$ ) distribution.) It is left to the reader to check that  $\nu_n$  converges weakly to  $\nu$ .

The approximating phase-type distributions are special cases of *mixtures*, and the phenomena that we observe in our example are typically also encountered in the mixture setting. Indeed, the large deviations of some mixtures are governed by nonconvex rate functions. This holds, in particular, for finite mixtures (in our case,  $\nu_n$  is a mixture of two distributions), of which the large deviations are readily analyzed. The infinite case is nontrivial; see [4], [7], [12], and [13]. Importantly, mixtures also arise naturally in connection with conditional probabilities; see, e.g. [17].

Consider the simulation of  $\nu_n(A)$ , where  $A := [0, \frac{1}{10}) \cup (10, \infty)$ . It is easy to see that  $\nu_n$  equals the distribution of  $(1/n) \sum_{i=1}^n X_i$  with probability  $\alpha$ , and equals that of  $(5/n) \sum_{i=1}^n X_i$  with probability  $1 - \alpha$ , where  $X_1, X_2, \dots$  have a standard exponential distribution. Using Cramér’s theorem, it is easy to see that  $\{\nu_n\}$  satisfies the LDP with the nonconvex good rate function

$$I(x) := \begin{cases} x - 1 - \log x & \text{if } x \in (0, \frac{5}{4} \log 5], \\ \frac{1}{5}x - 1 - \log \frac{1}{5}x & \text{if } x > \frac{5}{4} \log 5, \\ \infty & \text{otherwise.} \end{cases}$$

In order to avoid simulation, we could try to compute  $\nu_n(A)$  by calculating  $\nu_n^1(A)$  and  $\nu_n^5(A)$  numerically, where  $\nu_n^1$  is the law of  $(1/n) \sum_{i=1}^n X_i$  and  $\nu_n^5$  is the law of  $(5/n) \sum_{i=1}^n X_i$ . Indeed, both probabilities are readily expressed in terms of (incomplete) gamma functions. However, numerical problems arise even for moderate values of  $n$  since it is necessary to divide an incomplete gamma function by  $(n - 1)!$ .

Therefore, it is natural to attempt the estimation of  $\nu_n(A)$  using simulation techniques, for which there are several possibilities. Since  $\alpha$  is known, it suffices to simulate  $\nu_n^1(A)$  and  $\nu_n^5(A)$  separately. However,  $\nu_n^1(A)$  cannot be simulated efficiently by twisting exponentially; see [18] for related examples. To see that problems arise, we apply Corollary 1 to the simulation of  $\nu_n^1(A)$ , i.e. with the rate function  $I^1(x) := x - 1 - \log x$ . First note that  $\inf_{x \in A} I^1(x) = I^1(\frac{1}{10})$ ; we can readily check that the twist  $\xi = -9$  is the only candidate twist for asymptotic efficiency. However, this twist cannot be used for simulation, because the condition in Corollary 1 is violated: although

$$\inf_{x \in \mathbb{R}} I^1(x) + 9x = I^1(\frac{1}{10}) + \frac{9}{10},$$

we also have

$$\inf_{x \in A} I^1(x) - 9x < I^1(\frac{1}{10}) - \frac{9}{10}.$$

In the setting of this example, this can easily be overcome at some additional computational cost: we can simulate  $\nu_n^1((10, \infty))$  and  $\nu_n^1([0, \frac{1}{10}))$  separately. These probabilities and  $\nu_n^5(A)$  can be simulated efficiently using an exponential twist, meaning that a reliable estimate of  $\nu_n(A)$  can be found by simulating *three* different probabilities.

However, it is more efficient to take a direct approach in this case: the reader can easily check that the twist  $\xi = \frac{1}{10}$  is asymptotically efficient as a direct consequence of Corollary 1 (applied to the rate function  $I$ ). For the direct approach to be more efficient than the cutting approach, it is essential that we can easily sample from the  $\xi$ -twisted distribution.

The  $\xi$ -twisted measure  $\lambda_n$  is given, for Borel sets  $A$ , by

$$\lambda_n(A) = \frac{\alpha \int_A \exp[n\xi x] \nu_n^1(dx) + (1 - \alpha) \int_A \exp[n\xi x] \nu_n^5(dx)}{\alpha \int \exp[n\xi y] \nu_n^1(dy) + (1 - \alpha) \int \exp[n\xi y] \nu_n^5(dy)} \tag{6}$$

$$= \bar{\alpha}_n \frac{\int_A \exp[\frac{1}{10}nx] \nu_n^1(dx)}{(\frac{10}{9})^n} + (1 - \bar{\alpha}_n) \frac{\int_A \exp[\frac{1}{10}nx] \nu_n^5(dx)}{2^n}, \tag{7}$$

with

$$\bar{\alpha}_n = \frac{\alpha}{\alpha + (1 - \alpha)(\frac{9}{5})^n}.$$

Representation (7) shows how we can sample from  $\lambda_n$ : with probability  $\bar{\alpha}_n$ , we draw from an Erlang( $n, \frac{10}{9}n$ ) distribution and, with probability  $1 - \bar{\alpha}_n$ , we draw from an Erlang( $n, 10n$ ) distribution. Observe that  $\bar{\alpha}_n \rightarrow 0$ ; this is quite natural, as the mean of the twisted distribution then tends to  $10 = \arg \inf_{x \in A} I(x)$ . The likelihood ratio can be written as follows (cf. (6)):

$$\frac{d\lambda_n}{d\nu_n}(x) = \frac{\exp[\frac{1}{10}nx]}{\alpha(\frac{10}{9})^n + (1 - \alpha)2^n}.$$

Therefore, on the one hand, in the direct approach only one probability is simulated, instead of three as in the cutting approach. On the other hand, we must draw from the distribution  $(\bar{\alpha}_n, 1 - \bar{\alpha}_n)$ .

### 6. Discussion

In the case that any exponential twist for estimating  $\nu(A)$  is asymptotically inefficient, there are a number of alternatives. First of all, it may be possible to write the rare event  $A$  as a union of  $m < \infty$  disjoint rare events  $A_1, \dots, A_m$ , for which the probabilities can be estimated efficiently using an exponential twist; see also Section 5. The sum of these probabilities is then an asymptotically efficient estimator for  $\nu(A)$ . In many applications, however,  $A$  cannot be written as such a disjoint union. To overcome this, we can approximate  $\nu(A)$  by  $\nu(\bigcup_{i=1}^m A_i)$ , for suitably chosen  $A_1, \dots, A_m$ , and somehow bound the error, as in [5] and [11]. A variant of this approach is based on mixing relevant exponential twists; details can be found in [26]. In a hitting probability framework, Collamore [8] used related ideas to find an estimator that is arbitrarily ‘close’ to asymptotic efficiency.

Another way of dealing with asymptotically inefficient exponential twists is the recent adaptive approach to importance sampling described by Dupuis and Wang [14]. Although those authors illustrated this approach in an setting based on Cramér’s theorem, they claimed that it is useful in a more general setting. This dynamic exponential twisting contrasts with the approach taken in this paper, as we have considered a fixed exponential twist.

Although our definition of asymptotic efficiency is mathematically convenient, several other criteria for discriminating between estimators have been proposed. Notably, the amount of time (or work) required to generate one simulation replication was not taken into account in our definition of asymptotic efficiency. Glynn and Whitt [20] elaborated on a definition in which this is incorporated.

## References

- [1] ASMUSSEN, S. (1989). Risk theory in a Markovian environment. *Scand. Actuarial J.*, 69–100.
- [2] ASMUSSEN, S. (2003). *Applied Probability and Queues*, 2nd edn. Springer, New York.
- [3] ASMUSSEN, S. AND RUBINSTEIN, R. Y. (1995). Steady state rare events simulation in queueing models and its complexity properties. In *Advances in Queueing*, ed. J. Dshalalow, CRC, Boca Raton, FL, pp. 429–461.
- [4] BIGGINS, J. D. (2004). Large deviations for mixtures. *Electron. Commun. Probab.* **9**, 60–71.
- [5] BOOTS, N. K. AND MANDJES, M. (2002). Fast simulation of a queue fed by a superposition of many (heavy-tailed) sources. *Prob. Eng. Inf. Sci.* **16**, 205–232.
- [6] BUCKLEW, J. A., NEY, P. AND SADOWSKY, J. S. (1990). Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains. *J. Appl. Prob.* **27**, 44–59.
- [7] CHAGANTY, N. R. (1997). Large deviations for joint distributions and statistical applications. *Sankhyā A* **59**, 147–166.
- [8] COLLAMORE, J. F. (2002). Importance sampling techniques for the multidimensional ruin problem for general Markov additive sequences of random vectors. *Ann. Appl. Prob.* **12**, 382–421.
- [9] DEMBO, A. AND ZEITOUNI, O. (1998). *Large Deviations Techniques and Applications*, 2nd edn. Springer, New York.
- [10] DEUSCHEL, J.-D. AND STROOCK, D. W. (1989). *Large Deviations*. Academic Press, Boston, MA.
- [11] DIEKER, A. B. AND MANDJES, M. (2004). Fast simulation of overflow probabilities in a queue with Gaussian input. Tech. Rep. PNA-E0404, CWI.
- [12] DINWOODIE, I. H. AND ZABELL, S. L. (1992). Large deviations for exchangeable random vectors. *Ann. Prob.* **20**, 1147–1166.
- [13] DINWOODIE, I. H. AND ZABELL, S. L. (1993). Large deviations for sequences of mixtures. In *Statistics and Probability: A Raghu Raj Bahadur Festschrift*, eds J. K. Ghosh, S. K. Mitra, K. R. Parthasarathy and B. L. S. Prakasa Roa, John Wiley, New Delhi, pp. 171–190.
- [14] DUPUIS, P. AND WANG, H. (2004). Importance sampling, large deviations, and differential games. *Stoch. Stoch. Reports* **76**, 481–508.
- [15] FENG, J. AND KURTZ, T. (2000). Large deviations for stochastic processes. Preprint.
- [16] GLASSERMAN, P. AND KOU, S. (1995). Analysis of an importance sampling estimator for tandem queues. *ACM Trans. Model. Comput. Simul.* **4**, 22–42.
- [17] GLASSERMAN, P. AND LI, J. (2003). Importance sampling for portfolio credit risk. Preprint.
- [18] GLASSERMAN, P. AND WANG, Y. (1997). Counterexamples in importance sampling for large deviations probabilities. *Ann. Appl. Prob.* **7**, 731–746.
- [19] GLASSERMAN, P., HEIDELBERGER, P. AND SHAHABUDDIN, P. (1999). Asymptotically optimal importance sampling and stratification for pricing path-dependent options. *Math. Finance* **9**, 117–152.
- [20] GLYNN, P. W. AND WHITT, W. (1992). The asymptotic efficiency of simulation estimators. *Operat. Res.* **40**, 505–520.
- [21] HEIDELBERGER, P. (1995). Fast simulation of rare events in queueing and reliability models. *ACM Trans. Model. Comput. Simul.* **5**, 43–85.
- [22] LEHTONEN, T. AND NYRHINEN, H. (1992). On asymptotically efficient simulation of ruin probabilities in a Markovian environment. *Scand. Actuarial J.*, 60–75.
- [23] LEHTONEN, T. AND NYRHINEN, H. (1992). Simulating level-crossing probabilities by importance sampling. *Adv. Appl. Prob.* **24**, 858–874.
- [24] SADOWSKY, J. S. (1993). On the optimality and stability of exponential twisting in Monte Carlo estimation. *IEEE Trans. Inf. Theory* **39**, 119–128.
- [25] SADOWSKY, J. S. (1996). On Monte Carlo estimation of large deviations probabilities. *Ann. Appl. Prob.* **6**, 399–422.
- [26] SADOWSKY, J. S. AND BUCKLEW, J. A. (1990). On large deviations theory and asymptotically efficient Monte Carlo estimation. *IEEE Trans. Inf. Theory* **36**, 579–588.
- [27] SIEGMUND, D. (1976). Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.* **4**, 673–684.
- [28] SZECHTMAN, R. AND GLYNN, P. W. (2002). Rare-event simulation for infinite server queues. In *Proc. 34th Conf. Winter Simul. Exploring New Frontiers* (San Diego, CA, December 2002), eds E. Yücesan, C.-H. Chen, J. L. Snowdon and J. M. Charnes, IEEE, Piscataway, pp. 416–423.