

Large deviations for weighted empirical measures arising in importance sampling

Henrik Hult, Pierre Nyquist*

Department of Mathematics, KTH, 100 44 Stockholm, Sweden

Received 29 August 2014; received in revised form 1 June 2015; accepted 14 August 2015

Available online 21 August 2015

Abstract

In this paper the efficiency of an importance sampling algorithm is studied by means of large deviations for the associated weighted empirical measure. The main result, stated as a Laplace principle for these weighted empirical measures, can be viewed as an extension of Sanov's theorem. The main theorem is used to quantify the performance of an importance sampling algorithm over a collection of subsets of a given target set as well as quantile estimates. The analysis yields an estimate of the sample size needed to reach a desired precision and of the reduction in cost compared to standard Monte Carlo.

© 2015 Published by Elsevier B.V.

MSC: primary 60F10; 65C05; secondary 60G57; 68U20

Keywords: Large deviations; Empirical measures; Importance sampling; Monte Carlo

1. Introduction

Computational methods based on stochastic simulation – the collective term for simulating a physical system, involving random coefficients, on a computer – are fundamental in a wide range of applications, including the natural sciences, computer science, economics and finance, operations research, engineering sciences such as power grids, reliability, solid mechanics, etc. As the complexity of the mathematical models used in these subjects has increased, so has the

* Corresponding author.

E-mail addresses: hult@kth.se (H. Hult), pierre.nyquist@brown.edu (P. Nyquist).

need for fast and reliable computational methods. The aim of this paper is to introduce a new approach, based on the theory of large deviations for empirical measures, for analyzing the efficiency of stochastic simulation methods. In the present paper the emphasis is on importance sampling algorithms and the results can be viewed as a first step towards a new method for analyzing efficiency that can complement the existing variance-based approach.

The basic idea in stochastic simulation is to generate a population of particles that moves randomly according to the laws of the physical system. Each particle carries an individual weight, which may be updated during the simulation, and quantities of the underlying system are computed by averaging the particles' weights depending on their position. The canonical example is (standard) Monte Carlo simulation where all particles are independent and statistically identical, their weights being constant and equal.

Although the standard Monte Carlo method is widely used it is by no means universally applicable. One reason is that particles may wander off to irrelevant parts of the state space, leaving only a small fraction of relevant particles that contribute to the computational task at hand. Therefore, standard Monte Carlo may come with a computational cost that is too high for practical purposes. To remedy this, a control mechanism is needed that forces the particles to move to the relevant part of the state space, thereby reducing the computational cost. Such a control mechanism may come in different forms depending on the type of algorithm under consideration. In importance sampling, see e.g. [1], the control is the choice of sampling dynamics. In splitting algorithms [3] and genealogical particle systems [4,6] the control mechanism come, roughly speaking, in the form of a birth/death mechanism, which controls that important particles give birth to new particles and irrelevant particles are killed. For more on these and other related methods see [1,12] and the references therein.

Much of the theoretical work on the efficiency of stochastic simulation algorithms in general, and importance sampling algorithms in particular, is based on analyzing the variance of the resulting estimator. As long as an estimator is unbiased, variance is indeed the canonical measure of efficiency. However, when one considers situations in which the estimator is no longer unbiased, for example non-linear functionals of the underlying distribution, variance is no longer the obvious choice for determining performance and additional measures of efficiency can be useful.

Monte Carlo estimators are of the plug-in type where one first constructs an empirical measure associated with the distribution of interest and then plug this into whatever functional one is trying to estimate. A large deviation analysis for the underlying empirical measure will thus be useful in understanding the properties, such as rate of convergence, of the estimator. Therefore, much like for many other statistical estimators, a natural progression in the study of importance sampling algorithms and their performance is to consider the large deviation properties of the underlying weighted empirical measure. In the context of bootstrap estimators [13] derives such large deviation principles. Therein the sum of the weights is specified a priori and they are either deterministic or independent of the (possibly random) locations of the point masses in the empirical measure, making the setting quite different from that considered in this paper.

We propose in this paper that large deviation theory for the empirical measures arising in importance sampling can be used to define new performance criteria, based on the associated rate function, to complement the variance analysis. In addition to deriving the relevant large deviation result, we introduce one such criterion, which we refer to as evaluating performance over subsets. The intuition is that if a simulation algorithm is to approximate the original system in a certain region of the state space, then it should do so equally well over all subsets of that region that are not, in a sense, too small.

To avoid confusion it should be pointed out that techniques from sample-path large deviations have been studied thoroughly, to control the rate at which the variance decays as a large deviation parameter increases while keeping the number of particles fixed, in the context of designing efficient rare-event simulation algorithms; see for example [8,9] for results on the so-called subsolution approach. Our objective is fundamentally different. We want to complement the variance analysis by the rate function of a large deviation principle as the number of particles increases. For our purposes the appropriate framework is large deviations for empirical measures, in the spirit of Sanov's theorem [7, Theorem 2.2.1], rather than sample-path large deviations. The suggested approach is applicable to a wide range of simulation problems and is not intended exclusively for problems in rare-event simulation. Furthermore, the result is potentially useful for theoretical comparison of the performance of algorithms of different character, based on, say, importance sampling and interacting particle systems, by comparing the associated rate functions.

The contribution of the paper is two-fold. First, we prove a Laplace principle, in the space of finite measures equipped with the τ -topology, for the weighted empirical measure arising in importance sampling. This can be viewed as an extension of Sanov's theorem and under certain conditions the result follows from the latter and the contraction principle. Our proof of the general version, stated in Theorem 3.1, is based on the weak convergence approach to large deviations. It leads to a method for theoretical quantification of performance for importance sampling algorithms for non-linear functionals of the underlying distribution such as quantiles and other risk measures.

Second, we study the associated rate function and introduce a performance criteria for quantifying the performance of an importance sampling algorithm over a collection of subsets of a given target set as well as quantile estimates. A compelling feature of the large deviation approach is that it provides insights on the sample size needed to reach a desired accuracy as well as the reduction in computational cost for importance sampling compared to standard Monte Carlo. We do not explicitly address the optimal choice of sampling distribution, but rather the resulting performance given such a choice. The choice of sampling distribution may, for instance, be based on the subsolution approach [8,9].

The outline of the paper is as follows. In Section 2 an introduction to importance sampling is presented along with background on variance based efficiency analysis and large deviations for empirical measures. The main theoretical result, a Laplace principle for the weighted empirical measure of an importance sampling algorithm, is presented in Section 3. Applications of the large deviation result to efficiency analysis and design of importance sampling algorithms are given in Section 4. Most of the examples are well-studied elsewhere and are mainly intended to demonstrate the efficiency analysis using the rate function in contrast to the standard variance analysis. Section 5 contains the proof of the Laplace principle in Section 3.

2. Background and motivation

Let \mathcal{X} be a complete separable metric space equipped with its Borel σ -field $\mathcal{B}(\mathcal{X})$ and let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Unless otherwise stated subsets of \mathcal{X} under consideration are always assumed to be in $\mathcal{B}(\mathcal{X})$. Consider a random variable $X : \Omega \rightarrow \mathcal{X}$ with distribution $F = \mathbb{P} \circ X^{-1}$. Denote by \mathcal{M}_1 the space of probability measures on \mathcal{X} . The objective is to approximate F in a given region $A \in \mathcal{B}(\mathcal{X})$ or to compute $\Phi(F)$, where $\Phi : \mathcal{M}_1 \rightarrow \mathcal{R}$ is a given real-valued functional. Examples of functionals that may be of interest are expectations,

$$\Phi_f(F) = \int f dF, \quad \text{for some } f : \mathcal{X} \rightarrow \mathcal{R},$$

and, when X is real valued, quantiles

$$\Phi_q(F) = F^{-1}(q) = \inf\{x : F((x, \infty)) \leq q\}, \quad q \in (0, 1),$$

and L -statistics,

$$\Phi(F) = \int_0^1 \phi(q) F^{-1}(q) dq.$$

Only in exceptional cases is explicit computation of such functionals possible. When explicit computations are not possible a viable alternative is simulation. The standard method of simulation is Monte Carlo, in which the empirical measure

$$\mathbf{F}_n = \frac{1}{n} \sum_{k=1}^n \delta_{X_k},$$

is constructed from an independent sample X_1, \dots, X_n from F . Here δ_x denotes a unit point mass at x . The quantity $\Phi(F)$ is estimated by the plug-in estimator $\Phi(\mathbf{F}_n)$. Roughly speaking, if \mathbf{F}_n is a good approximation of F in a region that largely determines $\Phi(F)$, then $\Phi(\mathbf{F}_n)$ is likely to provide a good estimate.

2.1. Large deviation analysis for quantifying the performance of Monte Carlo algorithms

The precision of the estimator $\Phi(\mathbf{F}_n)$ will of course be affected by the sample size n . By the law of large numbers for empirical measures \mathbf{F}_n converges weakly to F with probability one, as $n \rightarrow \infty$, and an increased sample size n will thus improve the accuracy in this sense. Suppose $\Phi(\mathbf{F}_n)$ is an unbiased estimator for $\Phi(F)$. A natural requirement is to have a high probability of reaching a desired (good) precision:

$$\mathbb{P}(|\Phi(\mathbf{F}_n) - \Phi(F)| \leq \epsilon) \geq 1 - \delta,$$

for δ small and ϵ chosen in accordance with $\Phi(F)$ (e.g., so that one considers relative error). This is equivalent to the probability of an error greater than or equal to ϵ being smaller than or equal to δ . By Chebyshev's inequality this error probability is bounded from above according to

$$\mathbb{P}(|\Phi(\mathbf{F}_n) - \Phi(F)| \geq \epsilon) \leq \frac{1}{\epsilon^2} \text{Var}(\Phi(\mathbf{F}_n)).$$

Thus, for unbiased estimators $\Phi(\mathbf{F}_n)$ the sample size required to reach a desired precision is well-captured by $\text{Var}(\Phi(\mathbf{F}_n))$ and the performance of an estimator can be analyzed in terms of the variance; see [1] for different variance-based performance criteria. However, in the general case it may be that looking solely at the variance of the estimator is insufficient. One might use the mean-squared error $\mathbb{E}[(\Phi(\mathbf{F}_n) - \Phi(F))^2]$ instead but just as the variance this is not necessarily easy to analyze for a general Φ .

As an example of a situation in which alternative methods for analyzing the performance of Monte Carlo methods are needed one can consider the task of estimating a quantile. This is also one of the few cases of a more general estimation task, compared to estimating a probability, that has been addressed in the literature. Previous work in this area has primarily focused on methods involving central limit type results and the study of asymptotic properties of the estimators, a pioneering paper being [10]. The paper [11] reviews some recent results on Monte Carlo methods, specifically importance sampling, for estimating quantiles and expected shortfall

(denoted VaR and CVaR, respectively). As another example, [2] studies confidence intervals for quantile estimates arising from Monte Carlo methods. The authors use a Bahadur–Gosh representation for the estimators and derive associated central limit results.

An alternative way of quantifying the efficiency of the Monte Carlo estimator is through the theory of large deviations. To illustrate the point, let us consider the example of computing the expectation $F(f) = \int f dF$ for some $f : \mathcal{X} \mapsto \mathcal{R}$ by Monte Carlo simulation. An estimate of $F(f)$ is then given by

$$\mathbf{F}_n(f) = \frac{1}{n} \sum_{i=1}^n f(X_i), \quad (2.1)$$

where the X_i 's are independent with common distribution F .

Cramér's theorem states that if $\mathbb{E}[\exp\{\theta f(X)\}] < \infty$ for θ in a neighborhood of the origin, then $\mathbf{F}_n(f)$ satisfies the large deviation principle:

$$\begin{aligned} -I(A^\circ) &\leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(\mathbf{F}_n(f) \in A^\circ) \\ &\leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(\mathbf{F}_n(f) \in \bar{A}) \leq -I(\bar{A}), \end{aligned}$$

for all Borel sets A , where \bar{A} and A° denote the closure and interior of A , respectively, $I(A) = \inf_{x \in A} I(x)$, and the rate function I is given by

$$I(x) = \sup_{\theta} \{\theta x - \kappa(\theta)\},$$

where $\kappa(\theta) = \log \mathbb{E}[\exp\{\theta f(X)\}]$ is the logarithmic moment generating function [5, p. 26].

Suppose for the sake of illustration that, with probability at least $1 - \delta$, a high (relative) precision ϵ is desired in the estimate. That is, the sample size n must be selected sufficiently high that

$$\mathbb{P}(|\mathbf{F}_n(f) - F(f)| \geq \epsilon F(f)) \leq \delta.$$

Let A_ϵ denote the complement of the open ball of radius $\epsilon F(f)$ centered at $F(f)$,

$$A_\epsilon = \{x \in \mathcal{R} : |x - F(f)| \geq \epsilon F(f)\}.$$

By Cramér's theorem,

$$\limsup_n \frac{1}{n} \log \mathbb{P}(\mathbf{F}_n(f) \in A_\epsilon) \leq -I(A_\epsilon),$$

which implies that for large n , at least approximately,

$$\mathbb{P}(\mathbf{F}_n(f) \in A_\epsilon) \lesssim e^{-nI(A_\epsilon)}.$$

An upper bound δ on the error probability corresponds to

$$n \gtrsim \frac{1}{I(A_\epsilon)} (-\log \delta). \quad (2.2)$$

Roughly speaking the sample size must be proportional to the reciprocal of the rate for the error probability in order to obtain the desired performance.

The proposed method for analyzing performance of importance sampling algorithms is based on the assumption that δ is chosen sufficiently small. That is, that one is interested in having high precision with high probability. In the large deviation approximation $\mathbb{P}(\mathbf{F}_n(f) \in A_\epsilon)$, and thus also in (2.2), we have ignored any pre-exponential factors. This is standard when using large deviation results for approximating probabilities and for large n and/or small δ such factors play little to no role. If δ is not too small, so that less demand is put on the probability of high precision, then central limit type results could instead be more appropriate for quantifying the performance of an estimator.

Let us consider a more general example where we are interested in approximating the true distribution F over a region $A \in \mathcal{B}(\mathcal{X})$. Again the empirical measure \mathbf{F}_n resulting from Monte Carlo simulation, restricted to A , provides a viable approximation. In this context the large deviation principle for the sequence of empirical measures can be applied to quantify the performance of the Monte Carlo algorithm. Sanov's theorem [7, Theorem 2.2.1] states that the sequence $\{\mathbf{F}_n\}$ satisfies the large deviation principle on \mathcal{M}_1 with rate function given by the relative entropy $\mathcal{H}(\cdot | F)$, defined by

$$\mathcal{H}(G | F) = \int \log \frac{dG}{dF} dG, \quad G \in \mathcal{M}_1.$$

More precisely, for $A \in \mathcal{B}(\mathcal{M}_1)$

$$\begin{aligned} -I(A^\circ) &\leq \liminf_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(\mathbf{F}_n \in A^\circ) \\ &\leq \limsup_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{P}(\mathbf{F}_n \in \bar{A}) \leq -I(\bar{A}). \end{aligned}$$

Taking $A = U_G$ to be a small neighborhood of G one can interpret, for n large, $\exp\{-nI(G)\}$, where $I(G) = \mathcal{H}(G | F)$, as an approximation of the probability that the empirical measure looks like a typical sample from G . If it is undesirable that the empirical measure looks like a typical sample from a specific distribution G , say the objective is to have

$$\mathbb{P}(\mathbf{F}_n \in U_G) \leq \delta,$$

then, by the same reasoning as above, the sample size must be selected sufficiently large that

$$n \gtrsim \frac{1}{I(G)} (-\log \delta). \quad (2.3)$$

Notice that taking $A = A_\epsilon$ as defined above in Sanov's theorem one can recover (2.2).

In the context of a general functional Φ Sanov's theorem may also be applied. Either one would try to protect against a certain undesirable shape G of the empirical measure by selecting n as in (2.3) or by looking at error probabilities corresponding to sets of the form

$$A_\epsilon = \{G \in \mathcal{M}_1 : |\Phi(G) - \Phi(F)| \geq \epsilon \Phi(F)\}.$$

It may be pointed out that when considering a set such as A_ϵ the rate is given by

$$I(A_\epsilon) = \inf_{G \in A_\epsilon} I(G).$$

If the infimum is attained at G_ϵ , then limiting the probability of A_ϵ corresponds precisely to protecting against G_ϵ .

2.2. Importance sampling

The basic idea of importance sampling is to draw samples from a sampling distribution that is more likely to generate samples from the desired region, thus hopefully improving accuracy compared to standard Monte Carlo. Suppose that the goal is to evaluate $\Phi(F)$ for some functional Φ . For simplicity, start by considering the case of an expectation, $\Phi(F) = F(f) = \int f(x)F(dx)$ for an F -integrable, non-negative function $f : \mathcal{X} \rightarrow \mathcal{R}_+$. Let \tilde{F} be the chosen sampling distribution. We emphasize that for now we are not considering explicitly the question of how to choose/design \tilde{F} for a particular problem. Rather, the current work addresses how to analyze the performance of a suggested sampling distribution and comparing it to other choices of \tilde{F} and standard Monte Carlo. In Section 4 it is shown how such an analysis can indeed identify the best choice of sampling distribution within a given class, thus (at least partially) answering the design question as well. However, for now we think of \tilde{F} as given. Note that for \tilde{F} to be a feasible sampling distribution it must hold that $F \ll \tilde{F}$ on the support of f . If this holds, then the Radon–Nikodym derivative $dF/d\tilde{F}$ exists on $\{f > 0\}$ and it is possible to define the weight function

$$w(x) = \frac{dF}{d\tilde{F}}(x)I\{f(x) > 0\}. \quad (2.4)$$

Let $\tilde{X}_1, \dots, \tilde{X}_n$ be an independent sample from \tilde{F} . The weighted empirical measure corresponding to the importance sampling algorithm is

$$\tilde{\mathbf{F}}_n^w = \frac{1}{n} \sum_{k=1}^n w(\tilde{X}_k) \delta_{\tilde{X}_k}.$$

Note that in contrast to standard Monte Carlo $\tilde{\mathbf{F}}_n^w$ is typically not a probability measure. The importance sampling estimator of $F(f)$ is the plug-in estimator

$$\tilde{\mathbf{F}}_n^w(f) = \frac{1}{n} \sum_{i=1}^n w(\tilde{X}_i) f(\tilde{X}_i).$$

Let $\tilde{\mathbb{P}}$ and $\tilde{\mathbb{E}}$ denote the probability and expectation when $\tilde{X}_1, \tilde{X}_2, \dots$ are sampled from the sampling distribution \tilde{F} . If $\tilde{\mathbb{E}}[\exp\{\theta w(\tilde{X})f(\tilde{X})\}] < \infty$ for θ in a neighborhood of the origin, then Cramér's theorem implies that $\tilde{\mathbf{F}}_n^w(f)$ satisfies the large deviation principle with rate function

$$I^w(x) = \sup_{\theta} \{\theta x - \kappa^w(\theta)\},$$

where κ^w is the associated logarithmic moment generating function,

$$\kappa^w(\theta) = \log \tilde{\mathbb{E}}[\exp\{\theta w(\tilde{X})f(\tilde{X})\}].$$

Similar to the Monte Carlo illustration, suppose that a relative precision ϵ is desired in the estimate with probability at least $1 - \delta$. That is, the sample size n must be selected sufficiently high that

$$\tilde{\mathbb{P}}(|\tilde{\mathbf{F}}_n^w(f) - F(f)| \geq \epsilon F(f)) \leq \delta.$$

Just as before Cramér's theorem, with A_ϵ denoting the complement of the open ball of radius $\epsilon F(f)$ centered at $F(f)$, implies that the sample size must satisfy

$$n \gtrsim \frac{1}{I^w(A_\epsilon)} (-\log \delta).$$

The choice of the sampling distribution enters in the rate function through the weights $w = (dF/d\tilde{F})I\{f > 0\}$. The improvement over standard Monte Carlo can be quantified by comparing the rate function I corresponding to Monte Carlo and the rate function I^w corresponding to importance sampling. Furthermore, a good choice of the sampling distribution in the importance sampling algorithm is one that maximizes the rate $I^w(A_\epsilon)$.

To extend the analysis to more general functionals it is desirable to have an analogue of Sanov's theorem for the weighted empirical measures $\tilde{\mathbf{F}}_n^w$. In contrast to Monte Carlo one cannot expect that the weighted empirical measure $\tilde{\mathbf{F}}_n^w$ is a good approximation to F everywhere. Rather, the sampling distribution is selected to obtain good precision in the important part of the space. If the objective is to compute $\Phi(F)$ for some functional Φ , then it suffices that $\tilde{\mathbf{F}}_n^w$ approximates F in the region that largely determines $\Phi(F)$. For this purpose a non-negative measurable function $f : \mathcal{X} \rightarrow \mathcal{R}_+$ is introduced, called the *importance function*. Before proceeding some remarks are in place. First, the idea behind the importance function is to reflect the functional Φ involved in the estimation task at hand. It should therefore be selected on a case-by-case basis. An explicit example is to take Φ a quantile, $\Phi(F) = F^{-1}(p)$ for some p , and f the indicator function for the set $[b, \infty)$ for $b < F^{-1}(p)$: $f(x) = I\{x \in [b, \infty)\}$. Second, the importance function f also serves a technical purposes as it is used to ensure that the Radon–Nikodym derivative $dF/d\tilde{F}$ is well-defined on an appropriate part of the state space \mathcal{X} (as in the discussion leading up to (2.4)). Although not often explicitly addressed in the importance sampling literature, the principle behind the importance function is always present. For example, the zero-variance change of measure could not be treated without restricting the Radon–Nikodym derivative to only having to be defined on the sub- σ -algebra of interest. The use of a function f associated with Φ is a way to generalize this basic principle.

The rough statement that $\tilde{\mathbf{F}}_n^w$ is close to F in the important region is made precise by saying that the measure $\tilde{\mathbf{F}}_n^{wf}$ is close to F^f in the space $\mathcal{M} = \mathcal{M}(\mathcal{X})$ of finite measures, where

$$\tilde{\mathbf{F}}_n^{wf} = \frac{1}{n} \sum_{k=1}^n w(X_k) f(X_k) \delta_{X_k},$$

and F^f is the finite measure given by

$$F^f(g) = \int g(x) f(x) dF(x),$$

for each bounded measurable $g : \mathcal{X} \rightarrow \mathcal{R}$. In order for an efficiency analysis as the one described above to be feasible, the main objective is to establish an analogue of Sanov's theorem for the weighted empirical measures.

3. A Laplace principle for weighted empirical measures

In this section the main theoretical result of this paper is stated. It is an extension of Sanov's theorem to the weighted empirical measures arising in importance sampling, stated as a Laplace principle.

A sequence of random variables U_n taking values in a topological space \mathcal{U} is said to satisfy the *Laplace principle* on \mathcal{U} with rate function I if, for all bounded, continuous functions $h : \mathcal{U} \rightarrow \mathcal{R}$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{E}[e^{-nh(U_n)}] = - \inf_{u \in \mathcal{U}} \{h(u) + I(u)\}.$$

When \mathcal{U} is a Polish space, the Laplace principle is equivalent to the sequence satisfying the large deviation principle with the same rate function [7, Theorems 1.2.1, 1.2.3]. In the case of a general topological space, the relationship between the large deviation principle and the Laplace principle is given by Varadhan's lemma and Bryc's inverse, see [5] and the references therein.

Suppose that F and \tilde{F} are two given distributions on a complete, separable metric space \mathcal{X} . The space \mathcal{M} of finite measures on \mathcal{X} is equipped with the τ -topology: $\nu_n \xrightarrow{\tau} \nu$ if $\nu_n(g) \rightarrow \nu(g)$ for all bounded measurable $g : \mathcal{X} \rightarrow \mathcal{R}$. To avoid the subtle measurability issues of the τ -topology, see [7], we will have to work with $\mathcal{F}_{\mathcal{M}}$, the smallest σ -field on \mathcal{M} with respect to which the function mappings $\nu \mapsto \int g d\nu$ are measurable.

Let f be an importance function. That is, f is a non-negative F -integrable function characterizing the importance of different regions of \mathcal{X} . It is assumed that $F \ll \tilde{F}$ on the support of f and the weight function w is defined as in (2.4). Let \mathcal{M}_1 , the space of probability measures, be equipped with the τ -topology and introduce the set

$$\Gamma = \{G \in \mathcal{M}_1 : G(wf) < \infty\}.$$

Define the mapping Ψ from the subset $\Gamma \subset \mathcal{M}_1$ to \mathcal{M} as the mapping for which $\Psi(G; \cdot)$ is the finite measure given by

$$\Psi(G; g) = \int g(x) f(x) w(x) dG(x), \quad (3.1)$$

for each bounded measurable $g : \mathcal{X} \rightarrow \mathcal{R}$. A key observation is that $\tilde{\mathbf{F}}_n^{wf}(g) = \tilde{\mathbf{F}}_n(wfg)$, where $\tilde{\mathbf{F}}_n$ is the empirical measure obtained by sampling from \tilde{F} ,

$$\tilde{\mathbf{F}}_n = \frac{1}{n} \sum_{k=1}^n \delta_{\tilde{X}_k}.$$

Therefore $\tilde{\mathbf{F}}_n^{wf}(\cdot) = \Psi(\tilde{\mathbf{F}}_n; \cdot)$. Note also that $\tilde{\mathbf{F}}_n$ belongs to Γ with probability 1.

Let $\Delta \subset \mathcal{M}_1$ be the set

$$\Delta = \{G \in \mathcal{M}_1 : \mathcal{H}(G \mid \tilde{F}) < \infty\}. \quad (3.2)$$

We are now ready to introduce the rate function. Let $I : \mathcal{M} \mapsto [0, \infty]$ be the function defined by

$$I(\nu) = \inf\{\mathcal{H}(G \mid \tilde{F}) : G \in \Gamma \cap \Delta, \Psi(G) = \nu\}, \quad (3.3)$$

when such a G exists and $I(\nu) = \infty$ otherwise. Proposition 5.10 states that I has sequentially compact level sets. Our main result is the following.

Theorem 3.1 (*Laplace Principle for Weighted Empirical Measures*). *Let F and \tilde{F} be given as above and let f be an importance function. Suppose that*

- (i) *there exists a function $U : \mathcal{X} \rightarrow [0, \infty]$ such that $\int e^{U(x)} d\tilde{F}(x) < \infty$ and U has relatively compact level sets,*
- (ii) *$\int e^{\alpha w(x)f(x)} d\tilde{F}(x) < \infty$ for all $\alpha > 0$.*

The sequence $\{\tilde{\mathbf{F}}_n^{wf}\}$ of weighted empirical measures satisfies the Laplace principle on \mathcal{M} equipped with the τ -topology. For all bounded continuous functions $h : \mathcal{M} \rightarrow \mathcal{R}$, measurable on $(\mathcal{M}, \mathcal{F}_{\mathcal{M}})$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{E}[e^{-nh(\tilde{\mathbf{F}}_n^{wf})}] = - \inf_{v \in \mathcal{M}} \{h(v) + I(v)\}, \quad (3.4)$$

with the rate function I given by (3.3).

The proof is given in Section 5.

Remark 3.2. Condition (i) is a version of Condition 8.2.2 in [7], adapted to the case of independent and identically distributed variables. In the case of real-valued random variables it is a very mild assumption. Take, for example,

$$U(x) = \max\{\alpha \log(|x|), 0\},$$

for some $\alpha > 0$. Then U has relatively compact level sets and the condition of a finite expectation of $e^{U(X)}$ is in this case weaker than

$$\mathbb{E}[|X|^\alpha] < \infty,$$

for some $\alpha > 0$.

Remark 3.3. The right-hand side of (3.4) can be expressed as

$$\inf_{v \in \mathcal{M}} \{h(v) + I(v)\} = \inf_{G \in \Gamma \cap \Delta} \{h(\Psi(G)) + \mathcal{H}(G|\tilde{F})\}.$$

Although the expression on the left is the standard way to express the limit of the Laplace principle, the expression on the right better suits the weak convergence approach to large deviations that will be adopted throughout the proof.

Consider the special case when the function wf is bounded. Then, $\Psi : \Gamma \rightarrow \mathcal{M}$ is continuous when both spaces are equipped with the τ topology and Theorem 3.1 follows essentially from a standard application of the contraction principle, see, e.g., [7, Theorem 1.3.2]. The main difficulty is to show that the Laplace principle holds also in the general case.

4. Applications in performance analysis

In this section we propose one way in which the Laplace principle of Theorem 3.1 can be used to characterize the performance of an importance sampling algorithm. We outline a method for analyzing the performance over a collection of subsets of a target region A . If the sampling distribution is designed for a target set A , then one can expect to have good performance for subsets C of A that are not too small relative to A . To illustrate the performance analysis based on this criterion we consider some well-known examples. In addition to explicitly computing the rate function associated with Theorem 3.1 we obtain estimates on the reduction in computational cost when comparing standard Monte Carlo and importance sampling. We also discuss briefly the rare-event limit, when the target set has small probability. The section ends with a brief discussion on performance analysis for importance sampling algorithms designed for computing the quantile of a distribution.

4.1. Performance over a collection of subsets

In this section we are interested in the performance of importance sampling algorithms over a region $A \subset \mathcal{X}$, reflected in the importance function $f(x) = I\{x \in A\}$. The ideal is that the weighted empirical measure $\tilde{\mathbf{F}}_n^{wf}$ is close to F on all measurable subsets of A . For large n we can imagine that the weighted empirical measure looks like a typical sample from a measure ν , which is absolutely continuous with respect to F . The performance of the importance sampling algorithm is good if it is likely that $\tilde{\mathbf{F}}_n^{wf}$ looks like a typical sample from some ν belonging to a set of measures for which the likelihood ratio $d\nu/dF$ is close to 1 on A . For given $\epsilon > 0$ and $\delta > 0$ (where $\delta = \delta' F(A)$ for some $\delta' > 0$ is a reasonable choice), consider the sets

$$A_{\epsilon, \delta} = \{\nu \in \mathcal{M} : |d\nu/dF(x) - 1| \geq \epsilon \text{ for } x \in \text{some } C \subset A \text{ with } F(C) \geq \delta\},$$

$$A_{\epsilon, \delta}^+ = \{\nu \in \mathcal{M} : d\nu/dF(x) \geq 1 + \epsilon \text{ for } x \in \text{some } C \subset A \text{ with } F(C) \geq \delta\}, \quad (4.1)$$

$$A_{\epsilon, \delta}^- = \{\nu \in \mathcal{M} : d\nu/dF(x) \leq 1 - \epsilon \text{ for } x \in \text{some } C \subset A \text{ with } F(C) \geq \delta\}. \quad (4.2)$$

The rate $I(A_{\epsilon, \delta})$, with I as in [Theorem 3.1](#), can be used to evaluate the performance of the importance sampling algorithm. The interpretation is that $e^{-nI(A_{\epsilon, \delta})}$ is roughly the probability that $\tilde{\mathbf{F}}_n^{wf}$ provides an approximation of F with relative error greater than ϵ for some subset(s) $C \subset A$ that is not too small in the sense that $F(C) \geq \delta$. The sets $A_{\epsilon, \delta}^+$ and $A_{\epsilon, \delta}^-$ have similar interpretations for overestimation and underestimation, respectively. The sets $A_{\epsilon, \delta}^+$ and $A_{\epsilon, \delta}^-$ are somewhat easier to analyze than $A_{\epsilon, \delta}$, so for the sake of illustration they will be studied throughout the rest of this section. For brevity, in the examples that follows we will work exclusively with $A_{\epsilon, \delta}^+$.

If G is a probability measure such that $\nu = \Psi(G; \cdot)$ with Ψ as in [\(3.1\)](#), then $d\nu/dF = dG/d\tilde{F}$ on A and

$$I(A_{\epsilon, \delta}^+) = \inf \left\{ \mathcal{H}(G | \tilde{F}) : \frac{dG}{d\tilde{F}}(x) \geq 1 + \epsilon \text{ for } x \in \text{some } C \subset A, F(C) \geq \delta \right\}.$$

Denote by J_+ the set function given by

$$J_+(C) = \inf_{G \in \Gamma \cap \Delta} \left\{ \mathcal{H}(G | \tilde{F}) : \frac{dG}{d\tilde{F}}(x) \geq 1 + \epsilon \text{ for } x \in C \right\}.$$

The main theoretical result of this section is the following characterization of the large deviation rate $I(A_{\epsilon, \delta}^+)$.

Proposition 4.1. *For any $t \geq 0$, let*

$$C_t = \left\{ x \in A : \frac{dF}{d\tilde{F}}(x) \geq t \right\},$$

and denote by γ_ϵ^+ the function

$$\gamma_\epsilon^+(s) = (1 + \epsilon)s \log(1 + \epsilon) + (1 - (1 + \epsilon)s) \log\left(\frac{1 - (1 + \epsilon)s}{1 - s}\right).$$

For any $\delta > 0$, if there exists \tilde{t}_δ such that $F(C_{\tilde{t}_\delta}) = \delta$, then the large deviation rate for $A_{\epsilon, \delta}^+$ is given by

$$I(A_{\epsilon, \delta}^+) = \inf\{J_+(C) : C \subset A, F(C) \geq \delta\} = \gamma_\epsilon^+(\tilde{F}(C_{\tilde{t}_\delta})).$$

In general,

$$\sup\{J_+(C_t) : t \geq 0, F(C_t) \leq \delta\} \leq I(A_{\epsilon, \delta}^+) \leq \inf\{J_+(C_t) : t \geq 0, F(C_t) \geq \delta\}.$$

Before proceeding to a proof of [Proposition 4.1](#) we make the observation that γ_ϵ^+ is an increasing function. Therefore, with the reasoning in the previous section, a good choice of sampling distribution is one that makes $\tilde{F}(C_{\tilde{\delta}})$ large.

To prove [Proposition 4.1](#), that is compute $I(A_{\epsilon, \delta}^+)$, we first prove two intermediary results. Start by considering a fixed set $C \subset A$.

Lemma 4.2. *Given $C \subset A$ it holds that*

$$\inf\left\{\mathcal{H}(G \mid \tilde{F}) : \frac{dG}{d\tilde{F}}(x) \geq 1 + \epsilon \text{ for } x \in C\right\} = \gamma_\epsilon^+(\tilde{F}(C))$$

where the infimum is attained for the probability measure G^* with

$$\begin{aligned} \frac{dG^*}{d\tilde{F}}(x) &= 1 + \epsilon, \quad \text{for } x \in C, \\ \frac{dG^*}{d\tilde{F}}(x) &= \frac{1 - (1 + \epsilon)\tilde{F}(C)}{1 - \tilde{F}(C)}, \quad \text{for } x \in C^c. \end{aligned}$$

Proof. For any probability measure G with $\frac{dG}{d\tilde{F}}(x) \geq 1 + \epsilon$ on C , convexity of the function $\varphi(s) = s \log s$ and Jensen's inequality implies that

$$\begin{aligned} \mathcal{H}(G \mid \tilde{F}) &= \tilde{F}(C) \int_C \varphi\left(\frac{dG}{d\tilde{F}}(x)\right) \frac{d\tilde{F}(x)}{\tilde{F}(C)} + \tilde{F}(C^c) \int_{C^c} \varphi\left(\frac{dG}{d\tilde{F}}(x)\right) \frac{d\tilde{F}(x)}{\tilde{F}(C^c)} \\ &\geq \tilde{F}(C) \left(\frac{G(C)}{\tilde{F}(C)} \log \frac{G(C)}{\tilde{F}(C)} \right) + \tilde{F}(C^c) \left(\frac{G(C^c)}{\tilde{F}(C^c)} \log \frac{G(C^c)}{\tilde{F}(C^c)} \right) \\ &= G(C)(\log G(C) - \log \tilde{F}(C)) + (1 - G(C))[\log(1 - G(C)) - \log(1 - \tilde{F}(C))]. \end{aligned}$$

This is convex as a function of $G(C)$ and the minimizer is $G(C) = (1 + \epsilon)\tilde{F}(C)$. We conclude that the lower bound

$$\begin{aligned} &\inf\left\{\mathcal{H}(G \mid \tilde{F}) : \frac{dG}{d\tilde{F}}(x) \geq 1 + \epsilon \text{ for } x \in C\right\} \\ &\geq (1 + \epsilon)\tilde{F}(C) \log(1 + \epsilon) + (1 - (1 + \epsilon)\tilde{F}(C)) \log\left(\frac{1 - (1 + \epsilon)\tilde{F}(C)}{1 - \tilde{F}(C)}\right), \end{aligned}$$

holds. It is straightforward to check that the lower bound is attained by G^* . This completes the proof. \square

The rate $I(A_{\epsilon, \delta}^+)$ can be computed by minimizing the function J_+ over the collection of feasible sets.

Lemma 4.3. $I(A_{\epsilon, \delta}^+) = \inf\{J_+(C) : C \subset A, F(C) \geq \delta\}.$

Proof. For any $C^* \subset A$ such that $F(C^*) \geq \delta$ we have

$$\begin{aligned} I(A_{\epsilon, \delta}^+) &= \inf \left\{ \mathcal{H}(G \mid \tilde{F}) : \frac{dG}{d\tilde{F}}(x) \geq 1 + \epsilon \text{ for } x \in \text{some } C \subset A, F(C) \geq \delta \right\} \\ &\leq \inf \left\{ \mathcal{H}(G \mid \tilde{F}) : \frac{dG}{d\tilde{F}}(x) \geq 1 + \epsilon \text{ for } x \in C^* \right\} \\ &= J_+(C^*). \end{aligned}$$

Taking infimum over feasible sets C^* leads to

$$I(A_{\epsilon, \delta}^+) \leq \inf \{ J_+(C^*) : C^* \subset A, F(C^*) \geq \delta \}.$$

It remains to show the reverse inequality. For every $\eta > 0$ there exists a probability measure $G^* \in \Delta \cap \Gamma$ and a corresponding set $C^* \subset A$, with $F(C^*) \geq \delta$, such that

$$I(A_{\epsilon, \delta}^+) + \eta \geq \mathcal{H}(G^* \mid \tilde{F}) \geq J_+(C^*) \geq \inf \{ J_+(C) : C \subset A \text{ and } F(C) \geq \delta \}.$$

Since $\eta > 0$ is arbitrary the proof is complete. \square

To prove [Proposition 4.1](#) it only remains to characterize the minimizing set C of [Lemma 4.3](#) in terms of the likelihood ratio between F and \tilde{F} .

Proof of Proposition 4.1. The expression for $\mathcal{H}(G^* \mid \tilde{F})$ in [Lemma 4.2](#) is increasing in $\tilde{F}(C)$. Thus, minimizing $J_+(C)$ corresponds to taking infimum of $\tilde{F}(C)$. Consider the problem

$$\inf \{ \tilde{F}(C) : C \subset A, F(C) \geq \delta \}.$$

Let \mathcal{F}_A be the collection of all measurable functions $\alpha : A \mapsto [0, 1]$. Since indicator functions of subsets of A are included in \mathcal{F}_A ,

$$\inf \{ \tilde{F}(C) : C \subset A, F(C) \geq \delta \} \geq \inf \left\{ \int_A \alpha(x) d\tilde{F}(x) : \alpha \in \mathcal{F}_A, \int_A \alpha(x) dF(x) \geq \delta \right\},$$

with equality if the infimum on the right-hand side is attained for an indicator function.

Suppose that there exists \tilde{t}_δ with $F(C_{\tilde{t}_\delta}) = \delta$. We claim that $\alpha^*(x) = I\{x \in C_{\tilde{t}_\delta}\}$ is the solution to the problem on the right in the last display. To see this, take an arbitrary $\alpha \in \mathcal{F}_A$ such that $\int_A \alpha(x) dF(x) \geq \delta$.

$$\begin{aligned} \int_A (\alpha(x) - \alpha^*(x)) d\tilde{F}(x) &= \int_A (\alpha(x) - \alpha^*(x)) \frac{d\tilde{F}}{dF}(x) dF(x) \\ &= \int_{C_{\tilde{t}_\delta}} (\alpha(x) - 1) \frac{d\tilde{F}}{dF}(x) dF(x) + \int_{A \setminus C_{\tilde{t}_\delta}} \alpha(x) \frac{d\tilde{F}}{dF}(x) dF(x) \\ &\geq \int_{C_{\tilde{t}_\delta}} (\alpha(x) - 1) \frac{1}{\tilde{t}_\delta} dF(x) + \int_{A \setminus C_{\tilde{t}_\delta}} \alpha(x) \frac{1}{\tilde{t}_\delta} dF(x) \\ &= \frac{1}{\tilde{t}_\delta} \int_A \alpha(x) dF(x) - \frac{1}{\tilde{t}_\delta} \int_{C_{\tilde{t}_\delta}} dF(x) \\ &\geq 0, \end{aligned}$$

by the requirements on α and \tilde{t}_δ .

In the general case, it is obvious that

$$\inf \{ J_+(C) : C \subset A, F(C) \geq \delta \} \leq \inf \{ J_+(C_t) : t \text{ such that } F(C_t) \geq \delta \}.$$

Moreover, for any $t \geq 0$ such that $F(C_t) \leq \delta$ the same arguments as above yield

$$\int_A (\alpha(x) - I\{x \in C_t\}) d\tilde{F}(x) \geq \frac{1}{t} \int_A \alpha(x) dF(x) - \frac{1}{t} F(C_t) \geq 0,$$

for all $\alpha \in \mathcal{F}_A$ with $\int_A \alpha(x) dF(x) \geq \delta$. We conclude that

$$\inf \left\{ \int_A \alpha(x) d\tilde{F}(x) : \alpha \in \mathcal{F}_A, \int_A \alpha(x) dF(x) \geq \delta \right\} \geq \tilde{F}(C_t),$$

and it follows that

$$\inf \{J_+(C) : C \subset A, F(C) \geq \delta\} \geq J_+(C_t).$$

The proof is completed by taking supremum over t such that $F(C_t) \leq \delta$. \square

With the large deviation rate for $A_{\epsilon, \delta}^+$ explicitly computed we now outline the corresponding results for the set $A_{\epsilon, \delta}^-$ in (4.2). The results are obtained precisely as for $A_{\epsilon, \delta}^+$; J_- is the direct analogue of the set function J_+ .

Lemma 4.4. *Given $C \subset A$ it holds that*

$$\begin{aligned} & \inf \left\{ \mathcal{H}(G \mid \tilde{F}) : \frac{dG}{d\tilde{F}}(x) \leq 1 - \epsilon \text{ for } x \in C \right\} \\ &= (1 - \epsilon) \tilde{F}(C) \log(1 - \epsilon) + (1 - (1 - \epsilon) \tilde{F}(C)) \log \left(\frac{1 - (1 - \epsilon) \tilde{F}(C)}{1 - \tilde{F}(C)} \right), \end{aligned}$$

where the infimum is attained for the probability measure G_* with

$$\begin{aligned} \frac{dG_*}{d\tilde{F}}(x) &= 1 - \epsilon, \quad \text{for } x \in C, \\ \frac{dG_*}{d\tilde{F}}(x) &= \frac{1 - (1 - \epsilon) \tilde{F}(C)}{1 - \tilde{F}(C)}, \quad \text{for } x \in C^c. \end{aligned}$$

Lemma 4.5.

$$I(A_{\epsilon, \delta}^-) = \inf \{J_-(C) : C \subset A, F(C) \geq \delta\}.$$

Let γ_ϵ^- be the function

$$\gamma_\epsilon^-(s) = (1 - \epsilon)s \log(1 - \epsilon) + (1 - (1 - \epsilon)s) \log \left(\frac{1 - (1 - \epsilon)s}{1 - s} \right).$$

This is an increasing function in s and thus the optimal C is given precisely by the $C_{\tilde{t}_\delta}$ in Proposition 4.1. Thus, $I(A_{\epsilon, \delta}^-) = \gamma_\epsilon^-(\tilde{F}(C_{\tilde{t}_\delta}))$, whenever there exists \tilde{t}_δ such that $F(C_{\tilde{t}_\delta}) = \delta$.

Having obtained an explicit characterization of the rate function we now study some well-known examples in order to illustrate the performance analysis.

Example 4.6. Consider a standard Monte Carlo algorithm, i.e., $\tilde{F} = F$, and let A be a set with $F(A) = p$. Put

$$A_{\epsilon, \delta}^+ = \left\{ G \in \mathcal{M}_1 : \frac{dG}{dF}(x) \geq 1 + \epsilon \text{ for } x \in \text{some } C \subset A, F(C) \geq \delta \right\}.$$

For standard Monte Carlo, the rate function associated with Sanov's theorem is the relative entropy: $I^{MC}(G) = \mathcal{H}(G \mid F)$. The rate can be computed, just as in the general importance sampling case, as

$$I^{MC}(A_{\epsilon, \delta}^+) = \inf\{J_+^{MC}(C) : C \subset A, F(C) \geq \delta\},$$

with

$$J_+^{MC}(C) = \inf\left\{\mathcal{H}(G \mid F) : \frac{dG}{dF}(x) \geq 1 + \epsilon, x \in C\right\} = \gamma_\epsilon^+(F(C)).$$

Suppose there exists a set C such that $F(C) = \delta$. Then, since γ_ϵ^+ is increasing, we conclude that

$$I^{MC}(A_{\epsilon, \delta}^+) = \gamma_\epsilon^+(\delta),$$

and, by the reasoning leading to (2.2), the number of samples needed in order to obtain a specific error probability is proportional to the reciprocal of the rate $\gamma_\epsilon^+(\delta)$, with the proportionality constant depending only on the desired accuracy, i.e., bound on the error probability.

Let I^{IS} denote the rate function of Theorem 3.1 and suppose that in Proposition 4.1 such a \tilde{t}_δ exists. An importance sampling algorithm with sampling distribution \tilde{F} has rate

$$I^{IS}(A_{\epsilon, \delta}^+) = \gamma_\epsilon^+(\tilde{F}(C_{\tilde{t}_\delta})).$$

If the cost for generating one sample from \tilde{F} is c times the cost for generating one sample from F , then the reduction in computational cost is roughly

$$c \frac{n^{IS}}{n^{MC}} \approx c \frac{I^{MC}(A_{\epsilon, \delta}^+)}{I^{IS}(A_{\epsilon, \delta}^+)} = c \frac{\gamma_\epsilon^+(\delta)}{\gamma_\epsilon^+(\tilde{F}(C_{\tilde{t}_\delta}))}.$$

Example 4.7 (Light-Tailed Random Walk). For some $m \in \mathbb{N}$, let X_1, \dots, X_m be independent and identically distributed with finite moment generating function and let F denote the distribution the vector (X_1, \dots, X_m) . Consider the normalized light-tailed random walk S_m/m where $S_m = (X_1 + \dots + X_m)$. Take $A = \{x : \sum x_i \geq am\}$ to be the set of interest, so that $F(A) = \mathbb{P}(S_m \geq ma)$, and take the sampling distribution \tilde{F}_θ as an exponential change of measure:

$$\frac{dF}{d\tilde{F}_\theta}(x_1, \dots, x_m) = e^{m\left(\kappa(\theta) - \theta \frac{1}{m} \sum_{i=1}^m x_i\right)}.$$

Here κ is the logarithmic moment generating function of X_1 . Consider the set $A_{\epsilon, \delta}^+$ in (4.1). From Proposition 4.1 the rate $I(A_{\epsilon, \delta}^+)$ is given by $\gamma_\epsilon^+(\tilde{F}_\theta(C_{\tilde{t}_\delta}))$, where

$$C_{\tilde{t}_\delta} = \left\{x \in A : \frac{1}{m} \sum_{i=1}^m x_i \leq (\kappa(\theta) - (1/m) \log \tilde{t}_\delta)/\theta\right\}.$$

By the choice of \tilde{t}_δ ,

$$\begin{aligned} \tilde{F}_\theta(C_{\tilde{t}_\delta}) &= \mathbb{E}[e^{\theta S_m - m\kappa(\theta)} I\{a \leq S_m/m \leq (\kappa(\theta) - (1/m) \log \tilde{t}_\delta)/\theta\}] \\ &\geq e^{m(\theta a - \kappa(\theta))} \mathbb{E}[I\{a \leq S_m/m \leq (\kappa(\theta) - (1/m) \log \tilde{t}_\delta)/\theta\}] \\ &= e^{m(\theta a - \kappa(\theta))} F(C_{\tilde{t}_\delta}) \\ &= e^{m(\theta a - \kappa(\theta))} \delta. \end{aligned}$$

If the cost for each replication of the importance sampling algorithm is c times the cost for each replication of the standard Monte Carlo algorithm, then we conclude that the reduction in computational cost is given by

$$c \frac{\gamma_{\epsilon}^{+}(\delta)}{\gamma_{\epsilon}^{+}(\bar{F}(C_{i_{\delta}}))} \leq c \frac{\gamma_{\epsilon}^{+}(\delta)}{\gamma_{\epsilon}^{+}(e^{m(\theta_a - \kappa(\theta))} \delta)}.$$

A good choice of θ is the θ_a that maximizes $\theta a - \kappa(\theta)$. It holds that this θ_a is given by $\kappa'(\theta_a) = a$. We emphasize that in addition to suggesting the well known exponential change of measure with parameter θ_a , the large deviation analysis also provides this useful upper bound on the reduction in computational cost.

4.2. Applications in rare-event simulation

The efficiency analysis of the importance sampling algorithms presented so far is not targeted specifically to capture the performance of rare-event simulation algorithms. In this section we illustrate how a rare-event analysis can be performed, based on [Theorem 3.1](#) and the suggested performance criteria. The elementary examples presented in this section demonstrate that, based on [Theorem 3.1](#), one can obtain results that are consistent with variance analysis as well as estimates on the sample size required to reach a certain precision and the reduction in computational cost from using importance sampling.

We begin by analyzing standard Monte Carlo and the importance sampling algorithm for a light-tailed random walk with an emphasis on rare events.

Example 4.8 (Rare-Event Analysis for Standard Monte Carlo). Consider computing a probability $F(A) = p$ by standard Monte Carlo, as in [Example 4.6](#). Take $\delta = \delta' p$ for some $\delta' \in (0, 1)$. The performance of the algorithm can be captured by the rate $I^{MC}(A_{\epsilon, \delta' p}^{+}) = \gamma_{\epsilon}^{+}(\delta' p)$, where

$$\begin{aligned} \gamma_{\epsilon}^{+}(\delta' p) &= p\delta'(1 + \epsilon) \log(1 + \epsilon) + (1 - p\delta'(1 + \epsilon)) \log\left(\frac{1 - p\delta'(1 + \epsilon)}{1 - p\delta'}\right) \\ &= p\delta'[(1 + \epsilon) \log(1 + \epsilon) - \epsilon] + o(p^2). \end{aligned}$$

Thus, as $p \rightarrow 0$, the rate decays linearly with p and as a consequence the sample size needed for a given precision increases proportionally to $1/p$.

Example 4.9 (Rare-Event Analysis for a Light-Tailed Random Walk). In the random walk example, [Example 4.7](#), the performance of the importance sampling algorithm, with parameter $\theta = \theta_a$, is captured by $\gamma_{\epsilon}^{+}(e^{m(\theta_a a - \kappa(\theta_a))} \delta)$. Let $p_m = \mathbb{P}(S_m \geq ma)$ and take $\delta = \delta' p_m$. It follows that

$$\begin{aligned} \gamma_{\epsilon}^{+}(e^{m(\theta_a a - \kappa(\theta_a))} \delta' p_m) &= e^{m(\theta_a a - \kappa(\theta_a))} p_m \delta' [(1 + \epsilon) \log(1 + \epsilon) - \epsilon] \\ &\quad + o(e^{2m(\theta_a a - \kappa(\theta_a))} p_m^2). \end{aligned}$$

The reduction in computational cost for the importance sampler vs. the standard Monte Carlo algorithm is then bounded from above by

$$c \frac{\gamma_{\epsilon}^{+}(\delta' p_m)}{\gamma_{\epsilon}^{+}(e^{m(\theta_a a - \kappa(\theta_a))} \delta' p_m)} \sim c e^{-m(\theta_a a - \kappa(\theta_a))}, \quad \text{as } m \rightarrow \infty.$$

The conclusion is that the reduction in computational cost is exponential in m .

We end this section by demonstrating the performance of the zero-variance change of measure [1, p. 127]. This choice of sampling distribution is optimal for estimating a probability in the sense that the variance of the estimator is zero; it is often used as a reference point for designing efficient sampling algorithms.

Example 4.10 (Zero-Variance Change of Measure). Consider the probability $p = F(A)$, for some $A \subset \mathcal{X}$ and distribution F and take the importance function to be $I\{x \in A\}$. The zero-variance sampling distribution is the distribution \tilde{F} given by

$$\frac{d\tilde{F}}{dF}(x) = \frac{I\{x \in A\}}{p}.$$

Let $\delta' \in (0, 1)$ and $\delta = \delta'p$. The likelihood ratio is constant over the entire set A and $\tilde{F}(C) = F(C)p^{-1}$ for any $C \subset A$. That is, $\tilde{F}(C) = F(C \mid A)$, the conditional probability under F of C given A . It follows that

$$J_+(C) = \inf \left\{ \mathcal{H}(G \mid \tilde{F}) : \frac{dG}{d\tilde{F}}(x) \geq 1 + \epsilon, x \in C \right\} = \gamma_\epsilon^+(\tilde{F}(C)) = \gamma_\epsilon^+\left(\frac{F(C)}{p}\right),$$

and the rate

$$I(A_{\epsilon, \delta'p}^+) = \gamma_\epsilon^+(\delta').$$

In particular, the rate $I(A_{\epsilon, \delta'p}^+)$ is independent of p .

For $\delta' = 1$, the case of having a relative error ϵ on the estimate of p , the rate corresponding to the zero-variance change of measure is ∞ . This follows from the fact that no probability measure G which is absolutely continuous with respect to \tilde{F} on A , and thus must satisfy $G(A) = 1$, can give rise to such an error. In this particular case the rate (∞) can be obtained without using large deviation results. Since good performance corresponds to a large rate, the zero-variance change of measure is clearly optimal also from the large deviation perspective brought forward in this paper.

Remark 4.11. For state-dependent problems, such as the probability that a small-noise diffusion exits a domain before a final time T , say, it is possible to consider an analogy to the subsolution approach by Dupuis and Wang [8,9]. Taking $\delta = e^{-\delta_0/\epsilon}$ and $t = t_\epsilon = e^{t_0/\epsilon}$ a Hamilton–Jacobi equation for the exponential rate associated with $\tilde{F}_\theta(C_{t_\epsilon}^-)$ can be obtained and used for designing \tilde{F} . This approach to finding a good sampling distribution \tilde{F} will be further investigated elsewhere.

4.3. Performance analysis for computing quantiles

Let the underlying space \mathcal{X} be the real line and consider computing a quantile of a distribution F on \mathcal{R} . For $\alpha \in (0, 1)$ the α -quantile of a finite measure ν on \mathcal{R} is defined by

$$\Phi_\alpha(\nu) = \inf\{x : \nu((x, \infty)) \leq \alpha\}.$$

For $\epsilon > 0$, consider the set

$$A_\epsilon^+ = \{\nu \in \mathcal{M} : \Phi_\alpha(\nu) \geq (1 + \epsilon)\Phi_\alpha(F)\}.$$

Let \tilde{F} be the sampling distribution of an importance sampling algorithm for computing $\Phi_\alpha(F)$. Suppose that the importance function is $I\{x \in (a, \infty)\}$ where $a \leq \Phi_\alpha(F)$ and let I be the

rate function in [Theorem 3.1](#). The performance of the importance sampling algorithm can be quantified by $I(A_\epsilon^+)$. Of course one may also consider A_ϵ^- defined in the obvious way, but for this illustration we work exclusively with A_ϵ^+ .

Let us compute the rate $I(A_\epsilon^+)$. First note that, with $q_{\alpha,\epsilon} = (1 + \epsilon)\Phi_\alpha(F)$,

$$A_\epsilon^+ = \{v \in \mathcal{M} : \Phi_\alpha(v) \geq q_{\alpha,\epsilon}\} = \{v \in \mathcal{M} : v(q_{\alpha,\epsilon}, \infty) \geq \alpha\}$$

and the rate is therefore given by

$$\begin{aligned} I(A_\epsilon^+) &= \inf\{\mathcal{H}(G | \tilde{F}) : v = \Psi(G), v(q_{\alpha,\epsilon}, \infty) \geq \alpha\} \\ &= \inf\left\{\mathcal{H}(G | \tilde{F}) : \int_{\mathcal{R}} I\{x > q_{\alpha,\epsilon}\} w(x) dG(x) \geq \alpha\right\}. \end{aligned}$$

The infimum is attained at G_α^* given by

$$\frac{dG_\alpha^*}{d\tilde{F}}(x) = \frac{e^{\lambda k(x)}}{M(\lambda)},$$

where $k(x) = I\{x > q_{\alpha,\epsilon}\} w(x)$, $M(\lambda) = \int e^{\lambda k(x)} \tilde{F}(dx)$ and λ is given as the solution to

$$\alpha = \frac{\partial_\lambda M(\lambda)}{M(\lambda)}. \quad (4.3)$$

To see that the infimum is attained at G_α^* , note that by the variational formula for relative entropy [7, Proposition 4.5.1], for all $\lambda \geq 0$ and $G \in \mathcal{M}_1$ such that $\int k(x) dG(x) \geq \alpha$,

$$\mathcal{H}(G | \tilde{F}) \geq \lambda\alpha - \log M(\lambda),$$

and the inequality is satisfied with equality for G_α^* . We have just proved the following.

Proposition 4.12. $I(A_\epsilon^+) = \lambda\alpha - \log M(\lambda)$ where λ is determined by (4.3).

Example 4.13. For a standard Monte Carlo algorithm the rate $I(A_\epsilon^+)$ can be explicitly computed. Indeed, in this case $k(x) = I\{x \geq q_{\alpha,\epsilon}\}$ and

$$M(\lambda) = e^\lambda p_{\alpha,\epsilon} + 1 - p_{\alpha,\epsilon},$$

with $p_{\alpha,\epsilon} = F((q_{\alpha,\epsilon}, \infty)) \leq \alpha$. The equation for λ becomes

$$\alpha = \frac{\partial_\lambda M(\lambda)}{M(\lambda)} = \frac{e^\lambda p_{\alpha,\epsilon}}{e^\lambda p_{\alpha,\epsilon} + 1 - p_{\alpha,\epsilon}},$$

which leads to

$$\lambda = \log\left(\frac{\alpha(1 - p_{\alpha,\epsilon})}{(1 - \alpha)p_{\alpha,\epsilon}}\right)$$

and finally

$$I(A_\epsilon^+) = \lambda\alpha - \log M(\lambda) = \alpha \log\left(\frac{\alpha}{p_{\alpha,\epsilon}}\right) + (1 - \alpha) \log\left(\frac{1 - \alpha}{1 - p_{\alpha,\epsilon}}\right).$$

That is, if $\mathcal{H}(\alpha \mid p_{\alpha,\epsilon})$ refers to the relative entropy between two Bernoulli distributions with parameters α and $p_{\alpha,\epsilon}$, respectively, then

$$I(A_\epsilon^+) = \mathcal{H}(\alpha \mid p_{\alpha,\epsilon}).$$

For a general importance sampling algorithm the expression for $I(A_\epsilon^+)$ in [Proposition 4.12](#) has to be worked out on a case-by-case basis.

5. Proof of [Theorem 3.1](#)

In this section the proof of the Laplace principle for the weighted empirical measures of importance sampling is presented. The proof relies on the weak convergence approach developed by Dupuis and Ellis [\[7\]](#). The three main steps of the proof are:

- (1) Derive a representation formula for the pre-asymptotic expectation

$$W^n = -\frac{1}{n} \log \tilde{\mathbb{E}}[e^{-nh(\tilde{\mathbf{F}}_n^{wf})}].$$

This is achieved by formulating a stochastic control problem that has a minimal cost function equal to W^n . In the setting considered here the representation formula reads

$$W^n = \inf_{\{G_{n,j}\}} \tilde{\mathbb{E}} \left[\frac{1}{n} \sum_{j=0}^{n-1} \mathcal{H}(G_{n,j}(\cdot \mid \tilde{\mathbf{F}}_{n,j}) \mid \tilde{F}) + h(\tilde{\mathbf{F}}_n) \right], \quad (5.1)$$

where $\tilde{\mathbf{F}}_{n,j}$ is the controlled process (empirical measure), $\tilde{\mathbf{F}}_{n,j+1} = \tilde{\mathbf{F}}_{n,j} + \frac{1}{n} \delta_{\tilde{X}_{n,j}}$ and $\tilde{\mathbf{F}}_n = \frac{1}{n} \sum_{j=0}^{n-1} \delta_{\tilde{X}_{n,j}}$, obtained by sampling $\tilde{X}_{n,j}$ from the distribution (control) $G_{n,j}(\cdot \mid \tilde{\mathbf{F}}_{n,j})$.

- (2) The representation formula is used to prove the Laplace principle lower bound,

$$\liminf_n \frac{1}{n} \log \tilde{\mathbb{E}}[e^{-nh(\tilde{\mathbf{F}}_n^{wf})}] \geq - \inf_{G \in \Delta \cap \Gamma} \{h(\Psi(G)) + \mathcal{H}(G \mid \tilde{F})\}.$$

- (3) The most involved step is to use the representation formula to prove the Laplace principle upper bound,

$$\limsup_n \frac{1}{n} \log \tilde{\mathbb{E}}[e^{-nh(\tilde{\mathbf{F}}_n^{wf})}] \leq - \inf_{G \in \Delta \cap \Gamma} \{h(\Psi(G)) + \mathcal{H}(G \mid \tilde{F})\}.$$

5.1. Representation formula

In this section we show the representation formula [\(5.1\)](#) for the pre-limit expectation

$$W^n = -\frac{1}{n} \log \tilde{\mathbb{E}}[e^{-nh(\tilde{\mathbf{F}}_n^{wf})}].$$

This is achieved by considering a related stochastic control problem. For a more thorough discussion see [\[7, Section 2.3\]](#).

The difference between the current setting and that of Sanov's theorem is that only a subset $\Gamma = \{G \in \mathcal{M}_1 : G(wf) < \infty\}$ of the space of probability measures is under consideration. The proof of the representation formula W^n is therefore very similar to the standard case. Recall that $\tilde{\mathbf{F}}_n^{wf}(g) = \tilde{\mathbf{F}}_n(wfg)$ for each measurable g . In particular,

$$\tilde{\mathbb{E}}[\tilde{\mathbf{F}}_n^{wf}(I_{\mathcal{X}})] = \tilde{\mathbb{E}}[\tilde{\mathbf{F}}_n(wf)] = F(f) < \infty,$$

and it follows that $\tilde{\mathbf{F}}_n \in \Gamma$ with probability 1. Let $\Gamma_n = \Gamma$ and define, for $j = 0, \dots, n-1$, recursively the sets $\Gamma_j \subset \mathcal{M}_{j/n}$ by

$$\Gamma_j = \left\{ G \in \mathcal{M}_{j/n} : \tilde{F} \left(\left\{ y : G + \frac{1}{n} \delta_y \in \Gamma_{j+1} \right\} \right) = 1 \right\}.$$

That is, if $G \in \Gamma_j$, then sampling Y from \tilde{F} implies that $G + n^{-1} \delta_Y \in \Gamma_{j+1}$ with probability 1.

Let \mathcal{M}_0 be the set containing only the null measure and introduce the measurable mapping $W^n : \bigcup_{k=0}^n (\{k\} \times \mathcal{M}_{k/n}) \rightarrow \bar{\mathcal{R}} = [-\infty, \infty]$ by

$$W^n(j, G) = \begin{cases} -\frac{1}{n} \log \tilde{\mathbb{E}} \left[e^{-nh(\Psi(\tilde{\mathbf{F}}_n))} \mid \tilde{\mathbf{F}}_{n,j} = G \right], & \text{for } G \in \Gamma_j, \\ \infty, & \text{for } G \in \Gamma_j^c, \end{cases} \quad (5.2)$$

for $j = 0, \dots, n-1$ and

$$W^n(n, G) = \bar{h}(G), \quad (5.3)$$

where

$$\bar{h}(G) = \begin{cases} h(\Psi(G)), & \text{for } G \in \Gamma, \\ \infty, & \text{for } G \in \Gamma^c. \end{cases} \quad (5.4)$$

In particular, we set $W^n = W^n(0, 0)$. Note that in (5.2) $G \in \mathcal{M}_{j/n}(\mathcal{X})$ is a subprobability measure, and the $\tilde{\mathbf{F}}_{n,j} = (1/n) \sum_{i=0}^{j-1} \delta_{\tilde{X}_i}$ are the empirical subprobability measures obtained by sampling the \tilde{X}_i 's from \tilde{F} . By the same argument as in [7], the Markov chain $\tilde{\mathbf{F}}_{n,j}$ can be used to obtain the recursion formula

$$W^n(j, G) = -\frac{1}{n} \log \int e^{-nW^n(j+1, G + \frac{1}{n} \delta_x)} d\tilde{F}(x).$$

Since h is bounded and continuous the mapping W^n is measurable, bounded from below, and bounded from above on Γ . Together with the recursion formula above, Proposition 4.5.1 in [7] gives that $W^n(j, G)$ can be written as

$$W^n(j, G) = \inf_{\tilde{G} \in \Delta} \left\{ \frac{1}{n} \mathcal{H}(\tilde{G} \mid \tilde{F}) + \int W^n \left(j+1, G + \frac{1}{n} \delta_x \right) d\tilde{G}(x) \right\}, \quad (5.5)$$

for $j = 0, \dots, n-1$, where

$$\Delta = \{G \in \mathcal{M}_1 : \mathcal{H}(G \mid \tilde{F}) < \infty\}.$$

Moreover, the infimum is attained at $\tilde{G}_{n,j} \in \Delta$ defined by the Radon–Nikodym derivative

$$\frac{d\tilde{G}_{n,j}}{d\tilde{F}}(x) = \frac{e^{-W^n(j+1, G + \frac{1}{n} \delta_x)}}{\int e^{-W^n(j+1, G + \frac{1}{n} \delta_y)} d\tilde{F}(y)}. \quad (5.6)$$

To derive the representation formula for W^n , consider the following related stochastic control problem. For $n \in \mathbb{N}$ and $j \in \{0, 1, \dots, n\}$, let $G_{n,j}$ be a stochastic kernel on \mathcal{X} given $\mathcal{M}_{j/n}$. A

controlled process $\{\bar{\mathbf{F}}_{n,j}\}$ is defined by $\bar{\mathbf{F}}_{n,0} = \tilde{\mathbf{F}}_{n,0} = 0$ and

$$\bar{\mathbf{F}}_{n,j} = \frac{1}{n} \sum_{k=0}^{j-1} \delta_{\bar{X}_{n,k}}, \quad \bar{\mathbf{F}}_n = \bar{\mathbf{F}}_{n,n},$$

where the conditional distribution of $\bar{X}_{n,k}$ given $\bar{\mathbf{F}}_{n,0}, \bar{\mathbf{F}}_{n,1}, \dots, \bar{\mathbf{F}}_{n,k}$ is

$$\bar{\mathbb{P}}(\bar{X}_{n,k} \in dx \mid \bar{\mathbf{F}}_{n,0}, \bar{\mathbf{F}}_{n,1}, \dots, \bar{\mathbf{F}}_{n,k}) = G_{n,k}(dx \mid \bar{\mathbf{F}}_{n,k}).$$

All random variables and the corresponding (controlled) empirical subprobability measures are for all n defined on a common probability space $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{\mathbb{P}})$ which will be used throughout the paper. For these dynamics, $j \in \{0, 1, \dots, n\}$ and $G \in \mathcal{M}_{j/n}$, define the minimal cost functions

$$\bar{W}^n(j, G) = \inf_{\{G_{n,j}\}} \bar{\mathbb{E}} \left[\frac{1}{n} \sum_{k=j}^{n-1} \mathcal{H}(G_{n,k}(\cdot \mid \bar{\mathbf{F}}_{n,k}) \mid \tilde{F}) + \bar{h}(\bar{\mathbf{F}}_n) \mid \bar{\mathbf{F}}_{n,j} = G \right]. \quad (5.7)$$

For $j = 0$ and $G = 0$ we set

$$\bar{W}^n = \bar{W}^n(0, 0) = \inf_{\{G_{n,j}\}} \bar{\mathbb{E}} \left[\frac{1}{n} \sum_{k=0}^{n-1} \mathcal{H}(G_{n,k}(\cdot \mid \bar{\mathbf{F}}_{n,k}) \mid \tilde{F}) + \bar{h}(\bar{\mathbf{F}}_n) \right].$$

Proposition 5.1. *Let \bar{W}^n be given by (5.7) and W^n be the solution to (5.2) and (5.3). Then $W^n = \bar{W}^n$.*

Proof. We begin by proving that $\bar{W}^n(j, G) \geq W^n(j, G)$ using backwards induction on j . Fix a control sequence $\{G_{n,j}\}$ and let $\{\bar{\mathbf{F}}_{n,j}\}$ denote the associated controlled process. Let $\bar{I}_n = \Gamma$ and define sets \bar{I}_j associated with the control sequence $\{G_{n,j}\}$ by

$$\bar{I}_j = \left\{ G \in \mathcal{M}_{j/n} : G_{n,j} \left(\left\{ y : G + \frac{1}{n} \delta_y \in \bar{I}_{j+1} \right\} \mid G \right) = 1 \right\}, \quad j = 0, \dots, n-1.$$

The definition is such that if at time j the controlled process $\bar{\mathbf{F}}_{n,j}$ lies in \bar{I}_j , then by sampling from $G_{n,j}, G_{n,j+1}, \dots, G_{n,n-1}$ the controlled process $\bar{\mathbf{F}}_n$ will belong to Γ with probability 1.

Consider first the case $j = n$. Clearly,

$$\bar{W}^n(n, G) = \bar{h}(G) = W^n(n, G),$$

and the claim is trivial for this j .

Take $j = n-1$. Suppose that $G \in \bar{I}_{n-1}$, so that $G_{n,n-1}(\{y : G + \frac{1}{n} \delta_y \in \bar{I}_n\} \mid G) = 1$. Using (5.2), (5.3) and (5.5), we have

$$\begin{aligned} & \bar{\mathbb{E}} \left[\frac{1}{n} \mathcal{H}(G_{n,n-1}(\cdot \mid \bar{\mathbf{F}}_{n,n-1}) \mid \tilde{F}) + \bar{h}(\bar{\mathbf{F}}_n) \mid \bar{\mathbf{F}}_{n,n-1} = G \right] \\ &= \bar{\mathbb{E}} \left[\frac{1}{n} \mathcal{H}(G_{n,n-1}(\cdot \mid \bar{\mathbf{F}}_{n,n-1}) \mid \tilde{F}) + \bar{h}(\bar{\mathbf{F}}_n) + \int W^n(n, \bar{\mathbf{F}}_n) dG_{n,n-1} \right. \\ & \quad \left. - \int W^n(n, \bar{\mathbf{F}}_n) dG_{n,n-1} \mid \bar{\mathbf{F}}_{n,n-1} = G \right] \\ &\geq \bar{\mathbb{E}} \left[W^n(n-1, \bar{\mathbf{F}}_{n,n-1}) + \bar{h}(\bar{\mathbf{F}}_n) - \int W^n(n, \bar{\mathbf{F}}_n) dG_{n,n-1} \mid \bar{\mathbf{F}}_{n,n-1} = G \right] \\ &= W^n(n-1, G). \end{aligned}$$

If instead $G \in \bar{\Gamma}_{n-1}^c$, then $G_{n,n-1}(\{y : G + \frac{1}{n}\delta_y \in \Gamma_n\} \mid G) < 1$ and since \bar{h} is infinite on $\bar{\Gamma}_n^c$, this implies

$$\bar{\mathbb{E}}\left[\frac{1}{n}\mathcal{H}(G_{n,n-1}(\cdot \mid \bar{\mathbf{F}}_{n,n-1}) \mid \tilde{F}) + \bar{h}(\bar{\mathbf{F}}_n) \mid \bar{\mathbf{F}}_{n,n-1} = G\right] = \infty.$$

This shows that

$$\bar{\mathbb{E}}\left[\frac{1}{n}\mathcal{H}(G_{n,n-1}(\cdot \mid \bar{\mathbf{F}}_{n,n-1}) \mid \tilde{F}) + \bar{h}(\bar{\mathbf{F}}_n) \mid \bar{\mathbf{F}}_{n,n-1} = G\right] \geq W^n(n-1, G),$$

for any choice of G .

Proceeding similarly for $j = n-2, n-3, \dots, 0$ shows that

$$\bar{\mathbb{E}}\left[\frac{1}{n}\sum_{k=j}^{n-1}\mathcal{H}(G_{n,k}(\cdot \mid \bar{\mathbf{F}}_{n,k}) \mid \tilde{F}) + \bar{h}(\bar{\mathbf{F}}_n) \mid \bar{\mathbf{F}}_{n,j} = G\right] \geq W^n(j, G),$$

for all G and j . Taking infimum over all admissible control sequences $\{G_{n,j}\} \subset \Delta$ proves the inequality.

Next, the reverse inequality $\bar{W}^n(j, G) \leq W^n(j, G)$ is proved. Consider the control sequence $\{\tilde{G}_{n,j}\}$ defined by (5.6). For this particular sequence it holds that $\tilde{\Gamma}_j = \Gamma_j$ for all j .

The case $j = n$ was handled above and thus we start by considering $j = n-1$. If $G \in \Gamma_{n-1}$, then by the definition of $\tilde{G}_{n,n-1}$ and (5.5),

$$\begin{aligned} & \bar{\mathbb{E}}\left[\frac{1}{n}\mathcal{H}(\tilde{G}_{n,n-1}(\cdot \mid \bar{\mathbf{F}}_{n,n-1}) \mid \tilde{F}) + \bar{h}(\bar{\mathbf{F}}_n) \mid \bar{\mathbf{F}}_{n,n-1} = G\right] \\ &= \bar{\mathbb{E}}\left[\frac{1}{n}\mathcal{H}(\tilde{G}_{n,n-1}(\cdot \mid \bar{\mathbf{F}}_{n,n-1}) \mid \tilde{F}) + W^n(n, \bar{\mathbf{F}}_n) \mid \bar{\mathbf{F}}_{n,n-1} = G\right] \\ &= \bar{\mathbb{E}}\left[\frac{1}{n}\mathcal{H}(\tilde{G}_{n,n-1}(\cdot \mid \bar{\mathbf{F}}_{n,n-1}) \mid \tilde{F}) \mid \bar{\mathbf{F}}_{n,n-1} = G\right] + \bar{\mathbb{E}}[W^n(n, \bar{\mathbf{F}}_n) \mid \bar{\mathbf{F}}_{n,n-1} = G] \\ &= \frac{1}{n}\mathcal{H}(\tilde{G}_{n,n-1}(\cdot \mid G) \mid \tilde{F}) + \int W^n\left(n, G + \frac{1}{n}\delta_y\right) d\tilde{G}_{n,n-1}(y \mid G) \\ &= W^n(n-1, G). \end{aligned}$$

If instead $G \in \Gamma_{n-1}^c$, then $\tilde{G}_{n,n-1}(\{y : G + \frac{1}{n}\delta_y \in \Gamma\} \mid G) < 1$ and since \bar{h} is infinite on Γ^c we have

$$\bar{\mathbb{E}}\left[\frac{1}{n}\mathcal{H}(\tilde{G}_{n,n-1}(\cdot \mid \bar{\mathbf{F}}_{n,n-1}) \mid \tilde{F}) + \bar{h}(\bar{\mathbf{F}}_n) \mid \bar{\mathbf{F}}_{n,n-1} = G\right] = \infty = W^n(n-1, G).$$

This shows that

$$\bar{\mathbb{E}}\left[\frac{1}{n}\mathcal{H}(\tilde{G}_{n,n-1}(\cdot \mid \bar{\mathbf{F}}_{n,n-1}) \mid \tilde{F}) + \bar{h}(\bar{\mathbf{F}}_n) \mid \bar{\mathbf{F}}_{n,n-1} = G\right] = W^n(n-1, G),$$

for all G . Proceeding similarly for $j = n-2, \dots, 0$ shows that

$$\bar{\mathbb{E}}\left[\frac{1}{n}\sum_{k=j}^{n-1}\mathcal{H}(\tilde{G}_{n,k}(\cdot \mid \bar{\mathbf{F}}_{n,k}) \mid \tilde{F}) + \bar{h}(\bar{\mathbf{F}}_n) \mid \bar{\mathbf{F}}_{n,j} = G\right] = W^n(j, G),$$

for all j and G . Taking infimum over all admissible control sequences $\{G_{n,j}\}$ in Δ yields the desired inequality. This completes the proof. \square

5.2. Laplace principle lower bound

In this section the Laplace principle lower bound,

$$\liminf_n \frac{1}{n} \log \mathbb{E}[e^{-nh(\tilde{\mathbf{F}}_n^{wf})}] \geq - \inf_{G \in \Delta \cap \Gamma} \{h(\Psi(G)) + \mathcal{H}(G \mid \tilde{F})\}, \quad (5.8)$$

is proved. With W^n as in (5.2), the lower bound is equivalent to the upper bound

$$\limsup_n W^n \leq \inf_{G \in \Delta \cap \Gamma} \{h(\Psi(G)) + \mathcal{H}(G \mid \tilde{F})\}, \quad (5.9)$$

which is proved by using the representation formula (5.7) for W^n , derived in Proposition 5.1 is used.

The following strong law of large numbers will play a role in proving (5.9). Note that in Proposition 5.2 the function f is any measurable, non-negative F -integrable function, not necessarily any specific importance function. However, any importance function is of course an example of the type of function used in the proposition. Henceforth, f will again be used denote an (unspecified) importance function.

Proposition 5.2 (Strong Law of Large Numbers Under Importance Sampling). *Let f be non-negative, measurable and F -integrable. Let $\{X_j\}$ be independent and identically distributed with common distribution F . Let \mathbf{F}_n^f be the weighted measure in \mathcal{M} determined by*

$$\mathbf{F}_n^f(g) = \frac{1}{n} \sum_{j=0}^{n-1} f(X_j)g(X_j),$$

for each bounded measurable function g . Then, with probability 1,

$$\mathbf{F}_n^f \xrightarrow{\tau} F^f,$$

in \mathcal{M} .

A proof of the corresponding result for empirical measures, i.e., no weights, is found in [7] and the case of weighted measures can be treated in a similar way; we omit the proof.

Fix a probability measure $G \in \Delta \cap \Gamma$. Define the sequence of controls $\{G_{n,j}\}$ by $G_{n,j} = G$ for each $j = 0, 1, \dots, n-1$, so that in every step, the control does not depend on the controlled process. All the $\tilde{X}_{n,j}$'s are independent and identically distributed with common distribution G and the associated controlled process $\tilde{\mathbf{F}}_n$ belongs to Γ with probability 1. Using the representation formula for W^n , it follows that

$$W^n \leq \tilde{\mathbb{E}} \left[\frac{1}{n} \sum_{k=0}^{n-1} \mathcal{H}(G \mid \tilde{F}) + \bar{h}(\tilde{\mathbf{F}}_n) \right] = \mathcal{H}(G \mid \tilde{F}) + \tilde{\mathbb{E}}[h(\Psi(\tilde{\mathbf{F}}_n))].$$

The product wf is non-negative, measurable and G -integrable. By the choice G and the construction of the controlled process $\tilde{\mathbf{F}}_n$, Proposition 5.2 implies that with probability 1, $\Psi(\tilde{\mathbf{F}}_n; \cdot) \xrightarrow{\tau} \Psi(G; \cdot)$. By assumption h is bounded and continuous with respect to the τ -topology. Thus, $h(\Psi(\tilde{\mathbf{F}}_n)) \rightarrow h(\Psi(G))$ with probability 1. By the dominated convergence theorem,

$$\tilde{\mathbb{E}}[h(\Psi(\tilde{\mathbf{F}}_n))] \rightarrow h(\Psi(G)),$$

and we conclude that for any $G \in \Delta \cap \Gamma$

$$\limsup_n W^n \leq \mathcal{H}(G \mid \tilde{F}) + h(\Psi(G)).$$

Finally, taking infimum over G in $\Delta \cap \Gamma$ on the right-hand side proves the upper bound (5.9), and thus the Laplace principle lower bound.

5.3. Laplace principle upper bound

Just as for the lower bound, the Laplace principle upper bound

$$\limsup_n \frac{1}{n} \log \mathbb{E}[e^{-nh(\tilde{\mathbf{F}}_n^{wf})}] \leq - \inf_{G \in \Delta \cap \Gamma} [h(\Psi(G)) + \mathcal{H}(G \mid \tilde{F})], \quad (5.10)$$

can be stated as a lower limit for the minimal cost W^n ,

$$\liminf_n W^n \geq \inf_{G \in \Delta \cap \Gamma} [h(\Psi(G)) + \mathcal{H}(G \mid \tilde{F})]. \quad (5.11)$$

To prove (5.11) it is enough to show that every subsequence has a further subsequence that satisfies the lower limit. Therefore, we henceforth work with a fixed subsequence also denoted by W^n .

Since $\mathcal{H}(\cdot \mid \tilde{F})$ is a convex function [7, Proposition 1.4.3],

$$\frac{1}{n} \sum_{j=0}^{n-1} \mathcal{H}(G_{n,j}(\cdot \mid \tilde{\mathbf{F}}_{n,j}) \mid \tilde{F}) \geq \mathcal{H}\left(\frac{1}{n} \sum_{j=0}^{n-1} G_{n,j}(\cdot \mid \tilde{\mathbf{F}}_{n,j}) \mid \tilde{F}\right),$$

and together with the representation formula this gives the lower bound

$$W^n \geq \inf_{\{\tilde{G}_n\}} \bar{\mathbb{E}}[\mathcal{H}(\tilde{G}_n \mid \tilde{F}) + \bar{h}(\tilde{\mathbf{F}}_n)],$$

where \tilde{G}_n is defined as $\tilde{G}_n = (1/n) \sum_{j=0}^{n-1} G_{n,j}(\cdot \mid \tilde{\mathbf{F}}_{n,j})$. By optimality, for every $\epsilon > 0$, there exists a control sequence $\{G_{n,j}\}$ and associated \tilde{G}_n such that

$$W^n + \epsilon \geq \bar{\mathbb{E}}[\mathcal{H}(\tilde{G}_n \mid \tilde{F}) + \bar{h}(\tilde{\mathbf{F}}_n)]. \quad (5.12)$$

There is no restriction on assuming $G_{n,j} \in \Delta \cap \Gamma$ for each j and for the remainder of this section this assumption is made.

To show the Laplace principle upper bound it is now enough to prove the following result.

Proposition 5.3. *Every subsequence of $\{(\tilde{G}_n, \tilde{\mathbf{F}}_n)\}$ has a further subsequence, also denoted $\{(\tilde{G}_n, \tilde{\mathbf{F}}_n)\}$, such that $(\Psi(\tilde{G}_n), \Psi(\tilde{\mathbf{F}}_n)) \xrightarrow{\tau} (\Psi(\tilde{\mathbf{F}}), \Psi(\tilde{\mathbf{F}}))$ with probability 1 along this subsequence, and where $\tilde{\mathbf{F}}$ belongs to $\Delta \cap \Gamma$ with probability 1.*

Suppose that Proposition 5.3 holds. A proof of the Laplace principle upper bound follows from Fatou's lemma and the lower semi-continuity of the relative entropy mapping $G \mapsto \mathcal{H}(G \mid \tilde{F})$:

$$\begin{aligned} \epsilon + \liminf_n W^n &\geq \liminf_n \bar{\mathbb{E}}[\mathcal{H}(\tilde{G}_n \mid \tilde{F}) + \bar{h}(\tilde{\mathbf{F}}_n)] \\ &\geq \bar{\mathbb{E}}[\liminf_n \mathcal{H}(\tilde{G}_n \mid \tilde{F}) + \liminf_n h(\Psi(\tilde{\mathbf{F}}_n))] \end{aligned}$$

$$\begin{aligned} &\geq \bar{\mathbb{E}}[\mathcal{H}(\bar{\mathbf{F}} | \tilde{F}) + \bar{h}(\bar{\mathbf{F}})] \\ &\geq \inf_{G \in \Delta \cap \Gamma} \left\{ \mathcal{H}(G | \tilde{F}) + h(\Psi(G)) \right\}, \end{aligned}$$

where in the last step we used that $\bar{\mathbf{F}}$ is in $\Delta \cap \Gamma$ with probability 1.

Proposition 5.3 is proved by a series of lemmas. The idea is to first work with \mathcal{M}_1 equipped with the weak topology and show that $\{(\bar{G}_n, \bar{\mathbf{F}}_n)\}$ is tight in this topology. By showing that $\bar{\mathbb{E}}[\mathcal{H}(\bar{G}_n | \tilde{F})]$ is uniformly bounded (**Lemma 5.4**) the tightness of $\{(\bar{G}_n, \bar{\mathbf{F}}_n)\}$ is obtained by showing tightness of each of the marginals. Prohorov's theorem implies relative compactness and thus each subsequence has a subsubsequence converging to some random element $(\bar{G}, \bar{\mathbf{F}})$. **Lemma 5.6**, which corresponds to [7, Lemma 2.5.1], concludes that $\bar{G} = \bar{\mathbf{F}}$ w.p. 1, thus establishing that w.p. 1 for each subsequence,

$$(\bar{G}_n, \bar{\mathbf{F}}_n) \xrightarrow{w} (\bar{\mathbf{F}}, \bar{\mathbf{F}}),$$

along some further subsequence. Next, **Lemma 5.7** establishes that the convergences of the marginals \bar{G}_n and $\bar{\mathbf{F}}_n$ to $\bar{\mathbf{F}}$ are still valid when \mathcal{M}_1 is equipped with the τ -topology. The main ingredient of the proof is an approximation argument introduced in [7, Lemma 9.3.3] and **Lemma 5.7** is a version of that result in the simpler setting where the underlying random variables are independent and identically distributed.

Once the convergence in \mathcal{M}_1 equipped with the τ -topology is established w.p. 1, it remains to show that it is preserved under the mapping Ψ . **Lemma 5.8** proves that $\bar{\mathbf{F}}$ is in Γ w.p. 1, which implies that $\Psi(\bar{\mathbf{F}}; \cdot)$ is well-defined. The main additional difficulty is handled in **Lemma 5.9** where a truncation argument is used to prove that $\Psi(\bar{G}_n; \cdot)$ converges to $\Psi(\bar{\mathbf{F}}; \cdot)$ and $\Psi(\bar{\mathbf{F}}_n; \cdot)$ converges to $\Psi(\bar{\mathbf{F}}; \cdot)$ in the τ -topology on \mathcal{M} .

Lemma 5.4. *For a sequence $\{\bar{G}_n\}$ of control sequences such that (5.12) holds and $G_{n,j} \in \Gamma \cap \Delta$ for each j and n , it holds that*

$$\sup_n \bar{\mathbb{E}}[\mathcal{H}(\bar{G}_n | \tilde{F})] < \infty.$$

Proof. Since \bar{h} is bounded on Γ , it is possible to find a constant $M < \infty$ such that \bar{h} over Γ is bounded by M ; $\sup_{G \in \Gamma} |\bar{h}(G)| \leq M$. The choice of control sequence $\{G_{n,j}\}$ implies that $\bar{\mathbf{F}}_n \in \Gamma$ with probability 1 for all n . Therefore,

$$\begin{aligned} \sup_n \bar{\mathbb{E}}[\mathcal{H}(\bar{G}_n | \tilde{F})] &= \sup_n \bar{\mathbb{E}}[\mathcal{H}(\bar{G}_n | \tilde{F}) - M] + M \\ &\leq \sup_n \bar{\mathbb{E}}[\mathcal{H}(\bar{G}_n | \tilde{F}) + \bar{h}(\bar{\mathbf{F}}_n)] + M \\ &\leq \sup_n (W^n + \epsilon) + M \\ &= \sup_n W^n + \epsilon + M. \end{aligned}$$

For each n , $\bar{\mathbb{E}}[\mathcal{H}(\bar{G}_n | \tilde{F})] < \infty$ and by the proof of the Laplace principle lower bound $\limsup_n W^n < \infty$, which shows that $\sup_n \bar{\mathbb{E}}[\mathcal{H}(\bar{G}_n | \tilde{F})] < \infty$. \square

The uniform boundedness established in **Lemma 5.4** is used to prove the tightness of the sequence of admissible control measures.

Lemma 5.5. Under condition (i) of [Theorem 3.1](#), the sequence

$$\left\{ \frac{1}{n} \sum_{i=0}^{n-1} G_{n,j}(\cdot \mid \bar{\mathbf{F}}_n) \times \bar{\mathbf{F}}_n \right\} = \{(\bar{G}_n \times \bar{\mathbf{F}}_n)\},$$

of admissible control measures in $\mathcal{M}_1 \times \mathcal{M}_1$ is tight in the weak topology.

[Lemma 5.5](#) is a special case of Proposition 8.2.5 in [7] which establishes the result in the more general context of Markov chains. The proof is therefore omitted.

Having established the tightness of the distributions of $\{(\bar{G}_n, \bar{\mathbf{F}}_n)\}$, the next step is to extend this to almost sure convergence of subsequences, in the weak topology, and study the limit.

Lemma 5.6. Given any subsequence of $\{(\bar{G}_n, \bar{\mathbf{F}}_n)\}$, there exists a further subsequence that converges in distribution to some random variable $(\bar{G}, \bar{\mathbf{F}})$, where $\bar{G} = \bar{\mathbf{F}}$ a.s.

The result in [Lemma 5.6](#) is practically identical to part (b) of Lemma 2.5.1 in [7]. The only difference is that we must now appeal to the tightness proved in [Lemma 5.5](#), whereas in [7] the underlying space is assumed to be compact; we omit the proof.

The next step is to show that the weak convergence of subsequences actually implies convergence of \bar{G}_n and $\bar{\mathbf{F}}_n$ in the τ -topology. The result, [Lemma 5.7](#), is a version of Lemma 9.3.3 in [7] adapted to the case of independent and identically distributed random variables. Recall that we are already working with a specific subsequence (indexed by n) and on a probability space where the convergences $\bar{G}_n \xrightarrow{w} \bar{\mathbf{F}}$ and $\bar{\mathbf{F}}_n \xrightarrow{w} \bar{\mathbf{F}}$ both occur with probability 1. Henceforth, \mathcal{M}_1 will be equipped with the τ -topology.

Lemma 5.7. Under the conditions of [Lemmas 5.4–5.6](#), there exists some subsequence of $n \in \mathbb{N}$ such that $\bar{G}_n \xrightarrow{\tau} \bar{\mathbf{F}}$ and $\bar{\mathbf{F}}_n \xrightarrow{\tau} \bar{\mathbf{F}}$ w.p. 1 along this subsequence.

An important ingredient in the proof is the following inequality, which will appear frequently in what follows: For $a, b \geq 0$ and $\sigma \geq 1$,

$$ab \leq e^{a\sigma} + \frac{1}{\sigma}(b \log(b) - b + 1), \quad (5.13)$$

see [7] for a proof.

With the almost sure convergence in the τ -topology established the final results needed to prove [Proposition 5.3](#) are obtained in [Lemmas 5.8](#) and [5.9](#). We emphasize that this part of the proof is the main difference compared to Sanov's theorem and the essential ingredient needed to extend the standard result to [Theorem 3.1](#).

Lemma 5.8. Under the assumption $\int e^{\alpha w f} d\tilde{F} < \infty$ for any $\alpha > 0$, it holds that

$$\sup_n \bar{\mathbb{E}}[\bar{G}_n(wf)] < \infty, \quad \text{and} \quad \sup_n \bar{\mathbb{E}}[\bar{\mathbf{F}}_n(wf)] < \infty.$$

It follows that $\bar{\mathbf{F}} \in \Gamma$ w.p. 1.

Proof. [Lemma 5.4](#) shows that $\sup_n \bar{\mathbb{E}}[\mathcal{H}(\bar{G}_n \mid \tilde{F})] < \infty$. Hence, each \bar{G}_n has almost surely a well-defined Radon–Nikodym derivative w_n with respect to \tilde{F} and by definition

$$\mathcal{H}(\bar{G}_n \mid \tilde{F}) = \int w_n \log(w_n) d\tilde{F}.$$

Since w , f and w_n are all non-negative functions, the inequality (5.13) with $\sigma = 1$ gives

$$\begin{aligned}\bar{G}_n(wf) &= \int wf d\bar{G}_n = \int wf w_n d\tilde{F} \\ &\leq \int e^{wf} d\tilde{F} + \int (w_n \log w_n - w_n - 1) d\tilde{F} \\ &= \int e^{wf} d\tilde{F} + \int \log w_n d\bar{G}_n \\ &= \int e^{wf} d\tilde{F} + \mathcal{H}(\bar{G}_n | \tilde{F}).\end{aligned}$$

Thus,

$$\bar{\mathbb{E}}[\bar{G}_n(wf)] \leq \bar{\mathbb{E}}[\mathcal{H}(\bar{G}_n | \tilde{F})] + \int e^{wf} d\tilde{F},$$

and

$$\sup_n \bar{\mathbb{E}}[\bar{G}_n(wf)] \leq \sup_n \bar{\mathbb{E}}[\mathcal{H}(\bar{G}_n | \tilde{F})] + \int e^{wf} d\tilde{F}.$$

By the assumption and Lemma 5.4 it holds that

$$\sup_n \bar{\mathbb{E}}[\bar{G}_n(wf)] < \infty.$$

Lemma 5.7 proves that $\bar{G}_n \xrightarrow{\tau} \bar{\mathbf{F}}$ with probability 1. For $m \in \mathbf{N}$, $wf \wedge m$ is a bounded, measurable function and the τ -convergence implies that

$$\infty > \limsup_n \bar{\mathbb{E}}[\bar{G}_n(wf)] \geq \limsup_n \bar{\mathbb{E}}[\bar{G}_n(wf \wedge m)] = \bar{\mathbb{E}}[\bar{\mathbf{F}}(wf \wedge m)].$$

Taking \limsup as $m \uparrow \infty$ together with a repeated use of Fatou's lemma shows that $\bar{\mathbb{E}}[\bar{\mathbf{F}}(wf)] < \infty$. Since $\bar{\mathbf{F}}(wf)$ is a non-negative random variable it follows that $\bar{\mathbf{F}} \in \Gamma$ a.s.

The boundedness of $\sup_n \bar{\mathbb{E}}[\bar{\mathbf{F}}_n(wf)]$ is proved by repeatedly conditioning on the controlled process.

$$\begin{aligned}\bar{\mathbb{E}}[\bar{\mathbf{F}}_n(wf)] &= \bar{\mathbb{E}}\left[\frac{1}{n} \sum_{j=0}^{n-1} (wf)(\bar{X}_{n,j})\right] \\ &= \bar{\mathbb{E}}\left[\bar{\mathbb{E}}\left[\frac{1}{n} (wf)(\bar{X}_{n,n-1}) + \frac{1}{n} \sum_{j=0}^{n-2} (wf)(\bar{X}_{n,j}) \mid \bar{\mathbf{F}}_{n,n-1}\right]\right] \\ &= \bar{\mathbb{E}}\left[\frac{1}{n} \int (wf)(x) dG_{n,n-1}(x \mid \bar{\mathbf{F}}_{n,n-1}) + \bar{\mathbf{F}}_{n,n-1}(wf)\right] \\ &= \bar{\mathbb{E}}\left[\frac{1}{n} G_{n,n-1}(wf \mid \bar{\mathbf{F}}_{n,n-1})\right. \\ &\quad \left.+ \bar{\mathbb{E}}\left[\frac{1}{n} (wf)(\bar{X}_{n,n-2}) + \frac{1}{n} \sum_{j=0}^{n-3} (wf)(\bar{X}_{n,j}) \mid \bar{\mathbf{F}}_{n,n-2}\right]\right] \\ &= \bar{\mathbb{E}}\left[\frac{1}{n} G_{n,n-1}(wf \mid \bar{\mathbf{F}}_{n,n-1}) + \frac{1}{n} G_{n,n-2}(wf \mid \bar{\mathbf{F}}_{n,n-2}) + \bar{\mathbf{F}}_{n,n-2}(wf)\right].\end{aligned}$$

Proceeding like this one obtains

$$\begin{aligned}\mathbb{E}[\bar{\mathbf{F}}_n(wf)] &= \mathbb{E}\left[\frac{1}{n}G_{n,0}(wf \mid \bar{\mathbf{F}}_{n,0}) + \cdots + \frac{1}{n}G_{n,n-1}(wf \mid \bar{\mathbf{F}}_{n,n-1})\right] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{j=0}^{n-1}\int w(x)f(x)dG_{n,j}(x \mid \bar{\mathbf{F}}_{n,j})\right] \\ &= \mathbb{E}[\bar{G}_n(wf)].\end{aligned}$$

This completes the proof. \square

The final step for proving [Proposition 5.3](#) is to show that the almost sure convergence in the τ -topology on \mathcal{M}_1 implies almost sure convergence in the τ -topology on \mathcal{M} for the corresponding mapped measures $\Psi(\bar{G}_n; \cdot)$ and $\Psi(\bar{\mathbf{F}}_n; \cdot)$.

Lemma 5.9. *Along the subsequence for which the convergence in the τ -topology holds, the convergences*

$$\Psi(\bar{G}_n; \cdot) \xrightarrow{\tau} \Psi(\bar{\mathbf{F}}; \cdot), \quad \text{and} \quad \Psi(\bar{\mathbf{F}}_n; \cdot) \xrightarrow{\tau} \Psi(\bar{\mathbf{F}}; \cdot),$$

in \mathcal{M} hold w.p. 1.

Proof. Define for $G \in \mathcal{M}_1$ and $m \in \mathbf{N}$ a truncated version $\Psi_m(G; \cdot)$ of the mapping Ψ as the finite measure given by

$$\Psi_m(G; g) = \int (wf \wedge m)gdG,$$

for each bounded, measurable function g . The function $wf \wedge m$ is bounded and measurable and thus Ψ_m is continuous with respect to the τ -topology on \mathcal{M}_1 . Therefore,

$$\Psi_m(\bar{G}_n; g) = \int (wf \wedge m)gd\bar{G}_n \rightarrow \int (wf \wedge m)gd\bar{\mathbf{F}} = \Psi_m(\bar{\mathbf{F}}; g),$$

as $n \rightarrow \infty$ along the particular subsequence. That is, for any bounded, measurable function g it holds w.p. 1 that

$$\lim_{n \rightarrow \infty} \Psi_m(\bar{G}_n; g) = \Psi_m(\bar{\mathbf{F}}; g).$$

Moreover, by [Lemma 5.8](#) the function wf is integrable with respect to $\bar{\mathbf{F}}$ w.p.1. Since $(wf \wedge m)g \rightarrow wfg$ as $m \rightarrow \infty$, the dominated convergence theorem implies that w.p. 1,

$$\lim_{m \rightarrow \infty} \lim_{n \rightarrow \infty} \Psi_m(\bar{G}_n; g) = \Psi(\bar{\mathbf{F}}; g),$$

for every bounded, measurable function g . Therefore, the desired convergence in \mathcal{M} with the τ -topology will follow if the order of the limit operators can be interchanged, which holds if

$$\lim_{m \rightarrow \infty} \sup_n \left| \int g w f d\bar{G}_n - \int g (w f \wedge m) d\bar{G}_n \right| = 0. \quad (5.14)$$

The proof of [Lemma 5.7](#) involves an application of the Skorohod representation theorem, which results in the sequence $\{\mathcal{H}(\bar{G}_n \mid \bar{F})\}$ being bounded a.s. and (5.14) then follows by an application of the inequality (5.13); the argument is similar to Lemma 1.4.3(d) in [7].

The aim is now to show that for each bounded measurable function g , $\Psi(\bar{\mathbf{F}}_n; g) \rightarrow \Psi(\bar{\mathbf{F}}; g)$ in probability in a way such that for indicator functions $g = I_A$ we can appeal to the first

Borel–Cantelli lemma to get almost sure convergence of the entire measure along some subsequence. For this, take any $\epsilon > 0$ and consider the upper bound

$$\begin{aligned} & \bar{\mathbb{P}}\left(\left|\int wfgd\bar{\mathbf{F}}_n - \int wfgd\bar{\mathbf{G}}_n\right| \geq 3\epsilon\right) \\ & \leq \bar{\mathbb{P}}\left(\left|\int (wf \wedge m)gd\bar{\mathbf{F}}_n - \int wfgd\bar{\mathbf{F}}_n\right| \geq \epsilon\right) \end{aligned} \quad (5.15)$$

$$+ \bar{\mathbb{P}}\left(\left|\int (wf \wedge m)gd\bar{\mathbf{F}}_n - \int (wf \wedge m)gd\bar{\mathbf{G}}_n\right| \geq \epsilon\right) \quad (5.16)$$

$$+ \bar{\mathbb{P}}\left(\left|\int wfgd\bar{\mathbf{G}}_n - \int (wf \wedge m)gd\bar{\mathbf{G}}_n\right| \geq \epsilon\right), \quad (5.17)$$

which holds for any $m \geq 0$. For $n \in \mathbf{N}$ and $j \in \{0, 1, \dots, n-1\}$, define $\bar{\mathcal{F}}_{n,j}$ as the σ -algebra generated by the controlled process up to time j ,

$$\bar{\mathcal{F}}_{n,j} = \sigma(\bar{\mathbf{F}}_{n,0}, \bar{\mathbf{F}}_{n,1}, \dots, \bar{\mathbf{F}}_{n,j}).$$

Recall that

$$\bar{\mathbb{P}}(\bar{X}_{n,j} \in dy \mid \bar{\mathbf{F}}_{n,0}, \bar{\mathbf{F}}_{n,1}, \dots, \bar{\mathbf{F}}_{n,j}) = G_{n,j}(dy \mid \bar{\mathbf{F}}_{n,j}),$$

that is $G_{n,j}(\cdot \mid \bar{\mathbf{F}}_{n,j})$ is a regular conditional distribution for $\bar{X}_{n,j}$ given $\bar{\mathcal{F}}_{n,j}$. To take care of (5.15), condition on the $\bar{\mathcal{F}}_{n,j}$'s to relate expectation of integrals with respect to $\bar{\mathbf{F}}_{n,j}$ to integrals with respect to the control measures $G_{n,j}$,

$$\begin{aligned} & \bar{\mathbb{P}}\left(\left|\int (wf \wedge m)gd\bar{\mathbf{F}}_n - \int wfgd\bar{\mathbf{F}}_n\right| \geq \epsilon\right) \leq \frac{1}{\epsilon} \bar{\mathbb{E}}\left[\left|\int (wf \wedge m)gd\bar{\mathbf{F}}_n - \int wfgd\bar{\mathbf{F}}_n\right|\right] \\ & \leq \frac{\|g\|_\infty}{\epsilon} \bar{\mathbb{E}}\left[\int (wf - wf \wedge m)d\bar{\mathbf{F}}_n\right] \\ & = \frac{\|g\|_\infty}{\epsilon} \bar{\mathbb{E}}\left[\frac{1}{n} \sum_{j=0}^{n-1} (wf - wf \wedge m)(\bar{X}_{n,j})\right] \\ & = \frac{\|g\|_\infty}{\epsilon} \bar{\mathbb{E}}\left[\frac{1}{n} \sum_{j=0}^{n-1} \bar{\mathbb{E}}[(wf - wf \wedge m)(\bar{X}_{n,j}) \mid \bar{\mathcal{F}}_{n,j}]\right] \\ & = \frac{\|g\|_\infty}{\epsilon} \bar{\mathbb{E}}\left[\frac{1}{n} \sum_{j=0}^{n-1} \int (wf - wf \wedge m)dG_{n,j}\right] \\ & = \frac{\|g\|_\infty}{\epsilon} \bar{\mathbb{E}}\left[\int (wf - wf \wedge m)d\bar{\mathbf{G}}_n\right]. \end{aligned}$$

Here, $\|g\|_\infty$ denotes the sup-norm over \mathcal{X} .

Next, consider the term in (5.17). Analogous to the above,

$$\bar{\mathbb{P}}\left(\left|\int (wf \wedge m)gd\bar{\mathbf{G}}_n - \int wfgd\bar{\mathbf{G}}_n\right| \geq \epsilon\right) \leq \frac{\|g\|_\infty}{\epsilon} \bar{\mathbb{E}}\left[\int (wf - wf \wedge m)d\bar{\mathbf{G}}_n\right].$$

For any $\eta > 0$, define the set

$$A(m, \eta) = \{x \in \mathcal{X} : (wf)(x) - (wf \wedge m)(x) > \eta\},$$

and note that

$$\begin{aligned} \int (wf - wf \wedge m) d\bar{G}_n &= \int_{A(m, \eta)} (wf - wf \wedge m) d\bar{G}_n + \int_{\mathcal{X} \setminus A(m, \eta)} (wf - wf \wedge m) d\bar{G}_n \\ &\leq \int_{A(m, \eta)} (wf - wf \wedge m) d\bar{G}_n + \eta. \end{aligned}$$

Together with the inequality (5.13) this yields

$$\bar{\mathbb{E}} \left[\int (wf - wf \wedge m) d\bar{G}_n \right] \leq \int_{A(m, \eta)} e^{\sigma wf} d\tilde{F} + \sup_n \frac{1}{\sigma} \bar{\mathbb{E}}[\mathcal{H}(\bar{G}_n | \tilde{F})] + \eta.$$

Finally, consider the term in (5.16). To show that it converges to zero a martingale argument, similar to the proof of [7, Lemma 8.2.7], can be used. Since $G_{n,j}(\cdot | \bar{\mathbf{F}}_{n,j})$ is a regular conditional distribution of $\bar{X}_{n,j}$, for any bounded measurable function $g : \mathcal{X} \rightarrow \mathcal{R}$, w.p.1,

$$\begin{aligned} \bar{\mathbb{E}} &\left[(wf \wedge m)g(\bar{X}_{n,j}) - \int (wf \wedge m)gdG_{n,j}(x | \bar{\mathbf{F}}_{n,j}) | \bar{\mathcal{F}}_{n,j} \right] \\ &= \bar{\mathbb{E}}[(wf \wedge m)g(\bar{X}_{n,j}) | \bar{\mathcal{F}}_{n,j}] - G_{n,j}((wf \wedge m)g | \bar{\mathbf{F}}_{n,j}) \\ &= G_{n,j}((wf \wedge m)g | \bar{\mathbf{F}}_{n,j}) - G_{n,j}((wf \wedge m)g | \bar{\mathbf{F}}_{n,j}) = 0. \end{aligned}$$

Hence,

$$\{(wf \wedge m)g(\bar{X}_{n,j}) - G_{n,j}((wf \wedge m)g | \bar{\mathbf{F}}_{n,j})\}_{j=0,1,\dots,n-1},$$

is a martingale difference sequence with respect to $\{\bar{\mathcal{F}}_{n,j}\}$. Moreover, for any $\epsilon > 0$,

$$\begin{aligned} \bar{\mathbb{P}} &\left(\left| \int (wf \wedge m)gd\bar{\mathbf{F}}_n - \int (wf \wedge m)gd\bar{G}_n \right| \geq \epsilon \right) \\ &\leq \frac{1}{\epsilon^2} \bar{\mathbb{E}} \left[\left| \int (wf \wedge m)gd\bar{\mathbf{F}}_n - \int (wf \wedge m)gd\bar{G}_n \right|^2 \right] \\ &= \frac{1}{\epsilon^2} \bar{\mathbb{E}} \left[\frac{1}{n^2} \left(\sum_{j=0}^{n-1} ((wf \wedge m)g(\bar{X}_{n,j}) - G_{n,j}((wf \wedge m)g | \bar{\mathbf{F}}_{n,j})) \right)^2 \right] \\ &= \frac{1}{\epsilon^2} \bar{\mathbb{E}} \left[\frac{1}{n^2} \left(\sum_{j=0}^{n-1} ((wf \wedge m)g(\bar{X}_{n,j}) - G_{n,j}((wf \wedge m)g | \bar{\mathbf{F}}_{n,j}))^2 \right. \right. \\ &\quad \left. \left. + \sum_{i=1, j \neq i}^{n-1} ((wf \wedge m)g(\bar{X}_{n,i}) - G_{n,i}((wf \wedge m)g | \bar{\mathbf{F}}_{n,i})) \right. \right. \\ &\quad \left. \left. \times ((wf \wedge m)g(\bar{X}_{n,j}) - G_{n,j}((wf \wedge m)g | \bar{\mathbf{F}}_{n,j})) \right) \right]. \end{aligned}$$

The second term vanishes when conditioning on the σ -algebras $\{\bar{\mathcal{F}}_{n,j}\}$. For the first term inside the expectation an upper bound on each of the summands is

$$\left((wf \wedge m)g(\bar{X}_{n,j}) - G_{n,j}((wf \wedge m)g | \bar{\mathbf{F}}_{n,j}) \right)^2 \leq 4\|g\|_\infty^2 m^2.$$

It follows that

$$\frac{1}{\epsilon^2} \bar{\mathbb{E}} \left[\frac{1}{n^2} \left(\sum_{j=0}^{n-1} ((wf \wedge m)g(\bar{X}_{n,j}) - G_{n,j}((wf \wedge m)g | \bar{\mathbf{F}}_{n,j})) \right)^2 \right] \leq \frac{4\|g\|_\infty^2 m^2}{\epsilon^2 n}.$$

Combining the upper bounds for (5.15)–(5.17) yields the upper bound

$$\begin{aligned} \bar{\mathbb{P}}\left(|\Psi(\bar{\mathbf{F}}_n; wfg) - \Psi(\bar{G}_n; wfg)| \geq 3\epsilon\right) \\ \leq \frac{2\|g\|_\infty}{\epsilon} \left(\int_{A(m,\eta)} e^{\sigma wf} d\tilde{F} + \frac{1}{\sigma} \sup_n \bar{\mathbb{E}}[\mathcal{H}(\bar{G}_n | \tilde{F})] + \eta \right) + \frac{4\|g\|_\infty m^2}{\epsilon^2 n}, \end{aligned}$$

which converges to 0 if we send n, m and σ to ∞ in that order. To see this, note that for any $\eta > 0$, we can choose m large enough so that $\int_{A(m,\eta)} e^{\sigma wf} d\tilde{F}$ is arbitrarily small. Indeed, we have the upper bound

$$\tilde{F}(A(m, \eta)) \leq \frac{1}{\eta} \int_{A(m,\eta)} (wf - wf \wedge m) d\tilde{F} \leq \frac{1}{\eta} \int (wf - wf \wedge m) d\tilde{F},$$

and by the dominated convergence theorem, since $\tilde{F}(wf) < \infty$,

$$\int (wf - wf \wedge m) d\tilde{F} \rightarrow 0, \quad m \rightarrow \infty.$$

Hence, for any choice of η , $\tilde{F}(A(m, \eta)) \rightarrow 0$ as $m \rightarrow \infty$. The convergence to 0 for the upper bound now follows from the assumption $\int e^{\sigma wf} d\tilde{F} < \infty$, and Lemma 5.4.

To complete the proof we want to use the same argument as in Lemma 9.3.3 in [7] together with the convergence for $\Psi(\bar{G}_n; \cdot)$. For this it is not enough to have the above convergence in probability but one needs to be able to find a subsequence for which the convergence is fast enough for an application of the first Borel–Cantelli lemma. To illustrate, say that we want a subsequence n_k such that for $n \geq n_k$,

$$\bar{\mathbb{P}}\left(|\Psi(\bar{\mathbf{F}}_n; wfg) - \Psi(\bar{G}_n; wfg)| \geq 3\epsilon\right) \leq 2^{-k}.$$

Start by choosing $\eta \leq 2^{-k}\epsilon/8$. Next, pick a σ_k such that

$$\frac{1}{\sigma_k} \sup_n \bar{\mathbb{E}}[\mathcal{H}(\bar{G}_n | \tilde{F})] \leq \frac{2^{-k}\epsilon}{8},$$

which is again possible due to Lemma 5.4. Having chosen this σ_k , pick m_k sufficiently large so that

$$\int_{A(m_k, \eta)} e^{\sigma_k wf} d\tilde{F} \leq \frac{2^{-k}\epsilon}{8}.$$

Finally, pick n_k such that

$$\frac{m_k^2}{n_k} \leq \frac{2^{-k}\epsilon^2}{16}.$$

If g is an indicator function, the above will yield a sequence $\{n_k\}$ such that the probability of interest is smaller than 2^{-k} . The key to the argument used in Lemma 9.3.3 in [7] is that, since the underlying space \mathcal{X} is assumed complete and separable, the Borel σ -algebra on \mathcal{X} is generated by a countable collection of Borel sets. Together with the above this yields the desired result, namely that

$$\Psi(\bar{\mathbf{F}}_n; \cdot) \xrightarrow{\tau} \Psi(\mathbf{F}; \cdot),$$

in \mathcal{M} w.p. 1. The reader is referred to [7] for the details. \square

Lemmas 5.4–5.9 complete the proof of Proposition 5.3.

We end this section by proving that the function I in (3.3) has sequentially compact level sets in the τ -topology.

Proposition 5.10. *The function $I : \mathcal{M} \rightarrow [0, \infty]$ in (3.3) has sequentially compact level sets on \mathcal{M} equipped with the τ -topology.*

Proof. Let

$$C(K) = \{v \in \mathcal{M} : I(v) \leq K\},$$

for $K < \infty$. Take any sequence $\{v_n\} \subset C(K)$. Since $I(v_j) \leq K$ for each j , there exists a sequence $\{G_n\} \subset \Delta \cap \Gamma$ such that $\Psi(G_j) = v_j$. Moreover, it must hold that $\mathcal{H}(G_j \mid \tilde{F}) \leq K + \epsilon$ for every $\epsilon > 0$ and each j . Hence,

$$\sup_j \mathcal{H}(G_j \mid \tilde{F}) < \infty.$$

The relative entropy has compact level sets in the τ -topology [7, Proposition 9.3.6]. Therefore, there exists some subsequence, also indexed by n , and some G_* such that

$$G_n \xrightarrow{\tau} G_*,$$

and the corresponding (finite) measures $v_n = \Psi(G_n)$, $v_* = \Psi(G_*)$ are in the set $C(K)$. It remains to prove that $v_n \xrightarrow{\tau} v_*$. To this end, we note that by the same arguments as in Lemma 5.8 it holds that

$$\sup_n G_n(wf) < \infty,$$

and

$$G_*(wf) < \infty.$$

The conditions used to prove Lemma 5.9 are therefore satisfied and it follows in the same way that for every bounded measurable g ,

$$v_n(g) = \Psi(G_n; g) \rightarrow \Psi(G_*; g) = v_*(g)$$

as $n \rightarrow \infty$. Hence, $v_n \xrightarrow{\tau} v_*$ and the level set $C(K)$ is indeed sequentially compact in the τ -topology. \square

Acknowledgments

The authors would like to thank the two anonymous referees for their useful suggestions and remarks, which lead to an improvement of the presentation of the material.

References

- [1] S. Asmussen, P.W. Glynn, Stochastic Simulation: Algorithms and Analysis, in: Stochastic Modelling and Applied Probability, vol. 57, Springer, New York, 2007.
- [2] F. Chu, M.K. Nakayama, Confidence intervals for quantiles when applying variance-reduction techniques, ACM Trans. Model. Comput. Simul. 22 (2) (2012) 25. Art. 10.
- [3] T. Dean, P. Dupuis, Splitting for rare event simulation: a large deviation approach to design and analysis, Stochastic Process. Appl. 119 (2) (2009) 562–587.
- [4] P. Del Moral, J. Garnier, Genealogical particle analysis of rare events, Ann. Appl. Probab. 15 (4) (2005) 2496–2534.

- [5] A. Dembo, O. Zeitouni, *Large Deviations Techniques and Applications*, second ed., in: *Applications of Mathematics* (New York), vol. 38, Springer-Verlag, New York, 1998.
- [6] P. Dupuis, Y. Cai, Analysis of an interacting particle method for rare event estimation, *Queueing Syst.* 73 (2013) 345–406.
- [7] P. Dupuis, R.S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, in: *Wiley Series in Probability and Statistics: Probability and Statistics*, John Wiley & Sons Inc., New York, 1997.
- [8] P. Dupuis, H. Wang, Importance sampling, large deviations and differential games, *Stoch. Stoch. Rep.* 76 (6) (2004) 481–508.
- [9] P. Dupuis, H. Wang, Subsolutions of an Isaacs equation and efficient schemes for importance sampling, *Math. Oper. Res.* 32 (3) (2007) 723–757.
- [10] P.W. Glynn, Importance sampling for Monte Carlo estimation of quantiles, in: *Mathematical Methods in Stochastic Simulation and Experimental Design: Proceedings of the 2nd St. Petersburg Workshop on Simulation*, Publishing House of Saint Petersburg University, Saint Petersburg, 1996, pp. 180–185.
- [11] L.J. Hong, Z. Hu, G. Liu, Monte Carlo methods for value-at-risk and conditional value-at-risk: A review, *ACM Trans. Model. Comput. Simul.* 24 (4) (2014) 37. Art. 22.
- [12] G. Rubino, B. Tuffin (Eds.), *Rare Event Simulation Using Monte Carlo Methods*, John Wiley & Sons, Ltd., Chichester, 2009.
- [13] J. Trashorras, O. Wintenberger, Large deviations for bootstrapped empirical measures, *Bernoulli* 20 (4) (2014) 1845–1878.