Large deviations for weighted empirical measures
and processes arising in importance sampling

Pierre Nyquist
Abstract

This thesis consists of two papers related to large deviation results associated with importance sampling algorithms. As the need for efficient computational methods increases, so does the need for theoretical analysis of simulation algorithms. This thesis is mainly concerned with algorithms using importance sampling. Both papers make theoretical contributions to the development of a new approach for analyzing efficiency of importance sampling algorithms by means of large deviation theory.

In the first paper of the thesis, the efficiency of an importance sampling algorithm is studied using a large deviation result for the sequence of weighted empirical measures that represent the output of the algorithm. The main result is stated in terms of the Laplace principle for the weighted empirical measure arising in importance sampling and it can be viewed as a weighted version of Sanov’s theorem. This result 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 method of proof is the weak convergence approach to large deviations developed by Dupuis and Ellis.

The second paper studies moderate deviations of the empirical process analogue of the weighted empirical measure arising in importance sampling. Using moderate deviation results for empirical processes the moderate deviation principle is proved for weighted empirical processes that arise in importance sampling. This result can be thought of as the empirical process analogue of the main result of the first paper and the proof is established using standard techniques for empirical processes and Banach space valued random variables. The moderate deviation principle for the importance sampling estimator of the tail of a distribution follows as a corollary. From this, moderate deviation results are established for importance sampling estimators of two risk measures: The quantile process and Expected Shortfall. The results are proved using a delta method for large deviations established by Gao and Zhao (2011) together with more classical results from the theory of large deviations.

The thesis begins with an informal discussion of stochastic simulation, in particular importance sampling, followed by short mathematical introductions to large deviations and importance sampling.
Acknowledgements

I am immensely grateful to my advisor Henrik Hult. There are a lot of things that I could name here as reasons for my gratitude, but any attempt would come up short in describing just how privileged I consider myself to be for having him as my advisor. Henrik is a tremendous role model both as a researcher and as a teacher and I strive to follow the example that he sets.

I thank my co-advisor Filip Lindskog for continuing to challenge and influence the way I think about mathematics. Through his encouragement and the graduate courses he has taught, Filip has also made sure that my mathematical toolbox continues to grow.

In the life of a graduate student, the licentiate degree is merely a short stop on the road to obtaining the Ph.D. There is still a lot more work to be done and therefore I do not think that this is the time to give everyone the thanks they truly deserve. To those not mentioned here, rest assured that when the time comes I will express just how deep my gratitude is.

Finally, I must mention the one person to whom I owe so much. Mere words will never suffice but for now they will have to do: Hanna, thank you.
1 Introduction

For researchers in the natural sciences stochastic simulation has emerged as an indispensable tool. The most famous example, Monte Carlo simulation, was introduced by Metropolis and Ulam in [23] in 1949 but the underlying idea has been around for much longer than that. A prominent example of how stochastic simulation, specifically Monte Carlo methods, can come to use is given by recent findings at CERN, the European center for research in particle physics. Hardly anyone in all of science can have missed the announcement that was made in July 2012 regarding the Higgs boson. During the live press conference and lectures, as the researchers of the two teams that had come up with the findings were explaining the results, Monte Carlo was repeatedly mentioned as an integral part of the work. So, in what may possibly be one of the most profound and exciting scientific findings of the last couple of decades, stochastic simulation played an important role. This alone should provide enough incitement to continue study such methods and their theoretical properties.

This thesis deals with the intersection of the two topics large deviation theory and importance sampling. The question that has been motivating the work presented here is "Can the theory of large deviations be used for efficiency analysis of stochastic simulation in general, and importance sampling algorithms in particular?". Before we address this question in more detail, we give a short and non-formal introduction to stochastic simulation.

The term stochastic simulation encompasses all methods used for simulating a physical system, involving random effects, on a computer. As is indicated above, the application of such methods can be found in all of the natural sciences as well as in computer science, finance etc. In virtually all areas where probability is applied there is a use for stochastic simulation because the systems being modeled are often too complex to allow for explicit (analytical) calculations. Since there is such a vast interest in stochastic simulation from people from rather different scientific fields the development of these computational methods has gone in different directions. Within the applied communities the design and application of stochastic simulation algorithms is nowadays a large business and substantial resources are being put to use in this area. However, development of simulation algorithms is often done in an ad-hoc way building on intuition. Sometimes this results in great computational gains, but from a theoretical perspective such heuristic basis’ are not completely satisfying. Moreover, in the applied community extensive numerical experiments are often used as the measure of performance of an algorithm and such “proof” of efficiency, as it turns out, is unreliable. Indeed, numerous examples exist for which the standard heuristics, which may or may not have worked in similar settings, have failed and the numerical evidence falsely suggested convergence of the algorithm. Sometimes sophisticated algorithms based on some heuristic have poorer performance
than the more generic algorithms used throughout different scientific disciplines. Such issues prompts for a thorough theoretical analysis alongside the work of practitioners and for good communication and collaboration between disciplines.

When trying to simulate a physical system on a computer one has to define the dynamics, or laws, involved. The idea of stochastic simulation is then to generate particles that move randomly according to the dynamics of the system. Each particle involved carries an individual weight and to obtain estimators the the particles' weights are averaged depending on their locations. The canonical example of a stochastic simulation algorithm and the one of most interest for the work in this thesis is Monte Carlo. In standard Monte Carlo simulation the particles are independent and statistically identical and their weights are constant and equal. Suppose now that the quantity of interest can be represented as a high-dimensional integral. Let $X$ be a random variable with distribution $F$ and, for a function $f$, let $F(f) = \mathbb{E}[f(X)]$, the expected value of $f$ under the law $F$. The distribution $F$ is used to describe the dynamics governing the physical system and the information that we are interested in is represented by the integral $F(f)$.

In standard Monte Carlo, independent copies $X_1, \ldots, X_n$ of $X$ are generated and the empirical mean $F_n(f) = (1/n) \sum_{i=1}^n f(X_i)$ is used as an estimate of $F(f)$. By the law of large numbers, the estimate converges to the true value as $n \to \infty$. Since precision increases with number of particles, the computational cost of this algorithm is determined by the number of particles needed for a desired precision times the cost for generating the outcome of each particle.

Due to its relative simplicity standard Monte Carlo is used in a wide range of applications. Despite this, the method has some drawbacks that make it far from universally applicable. One such drawback is that the particles may wander off to parts of the state space that are irrelevant for the problem of interest. This leaves only a small fraction of the particles in a region that actually contributes to the computational task at hand. Therefore, standard Monte Carlo may need a so large number of particles that the computational burden for reaching a desired precision is too heavy for any practical use. A common problem where this becomes an issue is the estimation of rare-event probabilities.

To resolve the problem that particles can wander off, a control mechanism needs to be introduced that will force the particles to stay in the relevant region of the state space. Such a control can come in many different forms depending on the type of algorithm. In importance sampling, see e.g. [1] and Section 3 below, the control is a change of dynamics in the system. Rather than sampling particles according to the original dynamics, one uses a sampling dynamics which is chosen in a way which forces the particles into the relevant part of the state space. Such a change in dynamics will of course introduce an error and this is corrected for by assigning individual weights
that depend on the trajectory of each particle. Since an importance sampling algorithm is completely determined by the choice of sampling dynamics, finding one that is appropriate for a given problem is the main difficulty. This type of algorithms is the main object of study in this thesis.

Since there can be many suggestions for what simulation algorithm to use for a given problem, one would like to be able to pick what is in some sense the best one. Clearly, there is no need to switch from a simulation algorithm which has been used for a long time to a newer one unless the new outperforms the old. The need for efficiency analysis in stochastic simulation is therefore apparent.

Much of the theoretical analysis of efficiency of stochastic simulation algorithms is based on analyzing the variance of the resulting estimators. This is particularly true for Monte Carlo methods. The motivation is that the expectation operator is linear, making estimators of such quantities unbiased and thus variance is the canonical measure of variability. Since this covers a whole lot of estimation problems, for example that of estimating a probability, most of the effort related to efficiency analysis has been placed on variance. However, as the effort to improve simulation algorithms for more and more complex systems continues, the questions we are trying to answer using simulation are also becoming more and more intricate. That is, the quantities we try to estimate are not necessarily as, in some sense, simple as an expectation. Thus, estimators may no longer be unbiased or it may happen that more information about the estimator is needed than provided by variance analysis. This requires additional methods for measuring efficiency.

As a more precise formulation than the motivating question stated at the beginning, the aim of the project from which this thesis has emerged is to develop a systematic and unified approach to the construction of efficient Monte Carlo simulation algorithms. The specific problem on which the author has worked is the development of large deviation results related to the empirical measures and processes that arise in importance sampling. Previously, results on sample-path large deviations have been used extensively to design importance sampling algorithms. The aim of this project is fundamentally different in the way we propose to use large deviation results. Accompanying every large deviation result is a so-called rate function, a concept explained in detail in Section 2. Roughly speaking, the rate function associated with a system describes the rate of convergence of estimation errors. Therefore, the rate function can be used to quantify efficiency of a sampling algorithm and it provides a tool for comparing different algorithms to each other. The main contribution of the work presented in this thesis is that we establish the theoretical framework needed for this type of analysis to be possible.

The rest of the thesis consists of a more detailed introduction to the two topics large deviation theory (Section 2) and importance sampling (Section 3). In Sections 2 and 3 we also review, albeit briefly, some of the historical
development of the two topics. The introductory part of the thesis ends with summaries of two scientific papers in Section 4.

2 Large deviations

One of the main topics of research in modern probability is the theory of large deviations. Loosely speaking, it is a theory that describes the behavior of a stochastic system when it deviates from its expected behavior. First pioneered by Harald Cramér, the methods and results related to such questions were unified as a theory by Donsker and Varadhan in the west [9, 10, 11] and by Freidlin and Wentzell in the east [16, 28, 29, 30, 31]. The importance of the subject is exemplified by the awarding of the Able prize in 2007 to S.R.S. Varadhan “for his fundamental contributions to probability theory and in particular for creating a unified theory of large deviations.” Before stating explicitly the definition that underlies all of large deviation theory we begin with a motivating example. Variations on this example can be found in any standard book on large deviations, such as [7, 8, 26], and it gives some of the flavor of classical large deviation results and arguments. This example will also be used in Section 3.

Let $X_1, X_2, \ldots$ be a sequence of independent real-valued random variables with distribution $F$, with support on all of $\mathbb{R}$, and mean $E[X_1] = 0$. Let $S_n = \sum_{i=1}^{n} X_i$ and suppose that we are interested in the probability $P(S_n \geq na)$ for $a > 0$. By the law of large numbers, $S_n/n$ converges to 0 with probability 1 and thus $\{S_n \geq na\}$ is a rare event for $a > 0$ and $n$ large.

By Chebyshev’s inequality, for any $\theta > 0$,

$$P(S_n \geq na) = P(e^{\theta S_n} \geq e^{n\theta a}) \leq e^{-n\theta a} E[e^{\theta S_n}] = e^{-n\theta a} e^{n\kappa(\theta)},$$

were $\kappa(\theta)$ is the log-moment generating function evaluated in $\theta$,

$$\kappa(\theta) = \log E[e^{\theta X_1}].$$

It is assumed that $\kappa(\theta)$ exists for $\theta$ in a neighborhood of the origin. Since the above inequality holds for every $\theta > 0$, we have

$$P(S_n \geq na) \leq \exp\{-n \sup_{\theta > 0} \{\theta a - \kappa(\theta)\}\}.$$

This upper bound leads to

$$\limsup_{n \to \infty} \frac{1}{n} \log P(S_n \geq na) \leq - \sup_{\theta > 0} \{\theta a - \kappa(\theta)\}.$$

It can be shown that the restriction to positive $\theta$ can be removed.
An accompanying asymptotic lower bound for \( \mathbb{P}(S_n \geq na) \) can be derived using a change of measure. Denote by \( \tilde{F}_\theta \) the probability measure defined through the Radon-Nikodym derivative
\[
\frac{d\tilde{F}_\theta}{dF}(x) = \frac{e^{\theta x}}{e^{\kappa(\theta)}}, x \in \mathbb{R}.
\]
Let \( \tilde{F}_n^\theta \) denote the product measure of \( n \) copies of \( \tilde{F}_\theta \). Then, with \( x = (x_1, \ldots, x_n) \) and \( s_n = \sum_{i=1}^{n} x_i \),
\[
\mathbb{P}(S_n \geq na) = \int_{\mathbb{R}^n} I\{s_n \geq na\} F(dx_1) \ldots F(dx_n)
\]
\[
= \int_{\mathbb{R}^n} I\{s_n \geq na\} e^{n(\kappa(\theta) - \theta s_n)} \tilde{F}_\theta(dx_1) \ldots \tilde{F}_\theta(dx_n)
\]
\[
\geq e^{n\kappa(\theta)} \int_{\mathbb{R}^n} I\{s_n \geq na\} e^{-n\theta s_n} \tilde{F}_n^\theta(dx).
\]
Now, take \( \theta^* \) to be such that
\[
\sup_{\theta} \{\theta a - \kappa(\theta)\} = \theta^* a - \kappa(\theta^*).
\]
Then the above amounts to, with \( \varepsilon > 0 \),
\[
\mathbb{P}(S_n \geq na) \geq e^{n\kappa(\theta^*)} \int_{\mathbb{R}^n} I\{s_n \geq na\} e^{-n\theta^* s_n} \tilde{F}_n^\theta(dx)
\]
\[
\geq e^{n\kappa(\theta^*)} \int_{\mathbb{R}^n} I\{n(a + \varepsilon) \geq s_n \geq na\} e^{-n\theta^* s_n} \tilde{F}_n^\theta(dx)
\]
\[
\geq e^{-n(\theta^*(a + \varepsilon) - \kappa(\theta^*))} \int_{\mathbb{R}^n} I\{n(a + \varepsilon) \geq s_n \geq na\} \tilde{F}_n^\theta(dx)
\]
\[
= e^{-n(\theta^*(a + \varepsilon) - \kappa(\theta^*))} \mathbb{P}\left( n(a + \varepsilon) \geq \sum_{i=1}^{n} \tilde{X}_i \geq na \right),
\]
where the \( \tilde{X}_i \)'s are independent with common distribution \( \tilde{F}_\theta \). Notice that by the choice of \( \theta^* \), under the law \( \tilde{F}_\theta \), the random variables \( \{\tilde{X}_i\} \) have mean \( a \). Using the central limit theorem it is possible to show that a lower bound is given by, for \( n \) sufficiently large, \( \exp\{-n(\theta^* a - \kappa(\theta^*)) + o(n)\} \). Taking logarithm and sending \( n \) to infinity,
\[
\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n \geq na) \geq -(\theta^* a - \kappa(\theta^*)) = -\sup_{\theta} \{\theta a - \kappa(\theta)\}.
\]
Combining the upper and lower bounds,
\[
\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n \geq na) = -\sup_{\theta} \{\theta a - \kappa(\theta)\}.
\]
This characterizes deviations of the normalized random walk from the law of large numbers behavior.

The cornerstone of the theory of large deviations is the large deviation principle. Informally, a sequence of random variables satisfy the large deviation principle (LDP) if there is a function, known as the rate function, which determines the exponential decay of the probability of rare events. Formally, a sequence of random variables \( \{X_n\} \) on a topological space \( \mathcal{X} \) satisfies the LDP with speed \( \lambda_n^{-1}, \lambda_n \to \infty \), if there is a lower-semicontinuous function \( I : \mathcal{X} \mapsto [0, \infty] \) such that for any measurable set \( A \),

\[
- \inf_{x \in A^o} I(x) \leq \limsup_{n \to \infty} \frac{1}{\lambda_n} \mathbb{P}(X_n \in A^o) \leq \limsup_{n \to \infty} \frac{1}{\lambda_n} \mathbb{P}(X_n \in \bar{A}) \leq - \inf_{x \in \bar{A}} I(x),
\]

where \( A^o \) and \( \bar{A} \) denotes the interior and closure of \( A \), respectively. Unless otherwise mentioned, the speed is taken to be \( \lambda_n = n \).

One of the truly classical and perhaps best known results is Cramér’s theorem (see [6] and Chapter 2 in [7]) which deals with large deviations for the empirical mean of independent and identically distributed random variables (as in the example above). Cramér first showed the result for real-valued random variables but it has since been extended to a much more general setting; see [7], Chapter 4.

The earliest works on large deviations, including that of Donsker and Varadhan and Freidlin and Wentzell, were mostly based on change of measure techniques. The idea is that a lower bound for the probability of an event can be obtained by switching from the original measure to a probability measure which puts more mass (higher probability) on the event in question. This way, the lower bound of the LDP can be shown. The method is illustrated in the example above. There is a fundamental connection between this type of change of measure to prove large deviation results and the choice of importance sampling algorithm for similar problems. This is further discussed in Section 3.

In the present work we are interested in the sequence of empirical measures associated with a sequence of random variables. Sanov’s theorem, first proved in [25], gives the corresponding LDP. Here, \( \mathcal{X} \) is a topological space and \( \mathcal{M}_1 = \mathcal{M}_1(\mathcal{X}) \) is the space of probability measures on \( \mathcal{X} \).

**Theorem 2.1** (Sanov’s theorem). Let \( \{X_n\} \) be a sequence of independent random variables with common distribution \( F \) on \( \mathcal{X} \) and consider the sequence of empirical measures \( \{F_n\} \), where

\[
F_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(\cdot).
\]
The sequence \( \{F_n\} \) satisfies the LDP in \( M_1 \) with rate function \( I(\cdot) \) given by the relative entropy \( \mathcal{H}(\cdot | F) \),

\[
I(G) = \mathcal{H}(G | F) = \int_{\mathcal{X}} \log \left( \frac{dG}{dF} \right) dG, \ G \in M_1.
\]

Notice that Sanov’s theorem is a special case of the abstract version of Cramér’s theorem if \( M_1 \) is equipped with the weak topology, making it a complete, separable metric space.

An insight that has been useful for proving large deviation results is that many fundamental concepts and results in the theory of weak convergence, for example relative compactness, tightness, convergence of subsequences and the continuous mapping theorem, all have analogues in the theory of large deviations. This has lead to the development of an approach to proving the LDP using methods similar to those for proving weak convergence. An excellent treatment of the similarities between concepts in large deviations and weak convergence, and of large deviation results for stochastic processes, is [15].

The weak convergence approach to large deviations

Above we mentioned the analogies between the theory of large deviations and the theory of weak convergence of probability measures. There is yet another connection between large deviations and the methods of weak convergence, developed by Dupuis and Ellis and described in [12]. Their use of weak convergence methods stem from a connection between large deviations and control theory. For this, the concept of the Laplace principle plays a crucial role. A sequence \( \{X_n\} \) of random variables is said to satisfy the Laplace principle with rate function \( I \) if, for every bounded, continuous mapping \( h: \mathcal{X} \rightarrow \mathbb{R} \),

\[
\lim_{n \to \infty} \frac{1}{n} \log \mathbb{E}[e^{-nh(X_n)}] = -\inf_{x \in \mathcal{X}} \{h(x) + I(x)\}.
\]

For a general topological space \( \mathcal{X} \) the relationship between the LDP and the Laplace principle is given by Varadhan’s lemma and Bryc’s inverse; see [7] and the references therein. If \( \mathcal{X} \) is a complete, separable metric space the Laplace principle is equivalent to the LDP [12, Theorems 1.2.1, 1.2.3]. This equivalency between the LDP and the Laplace principle reveals an analogy between the theories of large deviations and weak convergence of probability measures. In the theory of weak convergence, one can look at either convergence of expectations of bounded, continuous functionals, or upper and lower limits of probabilities of closed and open sets (see the Prokhorov theorem, [2], Theorem 2.1). The Laplace principle and the LDP play similar roles in large deviation theory, the former being a statement about expectations
of bounded, continuous functionals and the latter one about probabilities of closed and open sets.

In [12] the authors show how it is possible for a large class of discrete time large deviation problems to represent the pre-limit quantity in the Laplace principle as the minimal cost function to a stochastic control problem. The following is an example of such a variational representation used in the proof of Sanov’s theorem ([12], Chapter 2). An analogous result is proved and discussed in detail in Paper 1. Let $X_1, X_2, \ldots$ be a sequence of independent random variables taking values in a complete, separable metric space $\mathcal{X}$ and having common distribution $F$. Denote the empirical measure by $F_n$. Then, for any bounded, continuous mapping $h : \mathcal{M}_1 \to \mathbb{R}$,

$$-\frac{1}{n} \log \mathbb{E}[e^{-nh(F_n)}] = \inf_{\{G_{n,j}\}} \mathbb{E} \left[ \frac{1}{n} \sum_{i=0}^{n-1} \mathcal{H}(G_{n,j} | F) + h(F_n) \right]$$

where the infimum is over admissible sequences of control (probability) measures $\{G_{n,j}\}$. Here, $F_n$ and $\mathbb{E}$ are used to denote a sequence of controlled empirical measures and expectation under the corresponding law. The exact definitions and meaning of these concepts are not important for now, they are thoroughly explained in Section 1.5 and Chapter 2 in [12] and (somewhat less) in Paper 1 of this thesis. The main thing to understand from the result is that rather than working directly with the expectations in the Laplace principle one can consider a corresponding stochastic control problem. Thus, the task of showing the Laplace principle is cast into the problem of showing convergence of the stochastic control problem This can be done using various methods from the well-developed theory of weak convergence of probability measures, motivating the name weak convergence approach to large deviations. The approach has proved successful in for problems difficult to handle using standard methods, particularly in infinite-dimensional settings. Of additional interest for the work in this thesis, it turns out that the variational representations used in this approach to proving large deviation results are useful for designing efficient importance sampling algorithms.

### Interpretation of the rate function

The rate function plays a central role in the application of large deviation theory to real-world problems. The most obvious interpretation of the rate function is that it provides the exponential rate at which probabilities of rare events decay. Indeed, suppose that a sequence $\{X_n\}$ of random variables satisfies the LDP with rate function $I$. Furthermore, assume that the set $A$ of interest is a continuity set for $I$: $I(A^c) = I(\bar{A})$, where $I(A) = \inf_{x \in A} I(x)$. Then,

$$\mathbb{P}(X_n \in A) \approx e^{-nI(A)}. \quad (2.1)$$
Similar statements are of course available, albeit less transparent and in terms of upper and lower bounds, without the assumption of $A$ being a continuity set for $I$. The interpretation (2.1) is the basis for the main application of large deviation theory, namely approximating the probability of rare events. In particular, a larger rate implies faster decay.

In addition to yielding exponential estimates of certain probabilities, the rate function has a more subtle interpretation. Namely, the element which gives the infimum of the rate function describes the way the event in question is most likely to happen. This is perhaps best illustrated using an example. Consider again the normalized random walk $S_n/n$ introduced above. The derivation shows that the LDP holds with rate function $I$ given by

$$\inf_{x \geq a} I(x) = \sup_{\theta} \{\theta a - \kappa(\theta)\}.$$ 

Moreover, the supremum is attained for $\theta^*$ such that under the distribution $\tilde{F}_{\theta^*}$ the random variable $\tilde{X}$ has mean $a$. The interpretation is as follows. If $S_n/n$ is to exceed the threshold $a$, the most likely way for this to happen is that each of the random variables involved takes a value $\geq a$. The rarer the event is, in this case the larger $a$ is, the more profound is this behavior. This is easily illustrated using a computer: Pick a threshold $a$ and simulate a large number of copies of $S_n/n$, keeping the ones that exceed $a$. What are the individual contributions of the random variables?

### 3 Importance sampling

Importance sampling is a method for performing Monte Carlo simulations that, if done correctly, will reduce the computational cost of standard Monte Carlo. It is particularly when the quantities being estimated are largely determined by rare events that importance sampling can give significant computational gains. The underlying idea is almost deceptively simple and is perhaps best illustrated using an elementary toy problem.

Let $X$ have a $N(0,1)$ distribution and say that we are interested in the probability $P(X > 3)$. Just from picturing the famous bell-shaped curve of the normal density one can conclude that this probability should be quite small. Indeed, it is approximately $1.35 \times 10^{-3}$. Suppose we did not know this and wanted use simulation to find the probability. First, notice that $p = P(X > 3) = E[I\{X > 3\}]$. Recalling the idea behind standard Monte Carlo, sampling $X_1, \ldots, X_n$ independently from $N(0,1)$ yields an estimate $\hat{p}_n = (1/n) \sum_{i=1}^n I\{X_i > 3\}$. However, only a few particles $X_i$ will be in the interval $(3, \infty)$ and a large number $n$ is required in order to get good precision of $\hat{p}_n$. Since $\hat{p}_n$ is unbiased, the number of samples $n$ needed to reach a desired precision is determined by the relative error. A reasonable precision would be that the estimate $\hat{p}_n$ has a relative error of 10%. For
standard Monte Carlo this amounts to

$$0.1 = \sqrt{\frac{\text{Var}(\hat{p}_n)}{p}} = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{p} - 1}.$$  

Using that we know $p$ (approximately) in this toy problem, an approximation of the number of samples needed to reach the desired precision is $n \approx 5.5 \times 10^5$. What if the random variables were instead sampled from the normal distribution with variance 1 but mean 3? Then, approximately half of the particles will be in the region of interest. Can the error introduced by sampling from the wrong distribution be accounted for? Let $\varphi(\cdot; \mu, \sigma)$ be the normal density with mean $\mu$ and variance $\sigma^2$. Since $\varphi(x; \mu, \sigma) > 0$ for all $x \in \mathbb{R}$,

$$E[I\{X > 3\}] = \int_{\mathbb{R}} I\{x > 3\} \varphi(x; 0, 1) \, dx = \int_{\mathbb{R}} I\{x > 3\} \frac{\varphi(x; 0, 1)}{\varphi(x; 3, 1)} \varphi(x; 3, 1) \, dx.$$  

Notice that the latter integration is with respect to the $N(3,1)$ distribution. Setting $w(x) = \varphi(x; 0, 1)/\varphi(x; 3, 1)$,

$$p = E[I\{X > 3\}] = E[I\{Z > 3\} w(Z)],$$

where $Z$ is distributed according to $N(3,1)$. Hence, it is possible to sample $Z_1, \ldots, Z_n$ from $N(3,1)$ and use $\tilde{p}_n = (1/n) \sum_{i=1}^n I\{Z_i > 3\} w(Z_i)$ to estimate $p$. In the case of the normal distribution, since we have approximations for all tail probabilities, we can get an idea of the computational gain compared to standard Monte Carlo. Similar to the above analysis for $\hat{p}_n$, consider the relative error of $\tilde{p}_n$ being 10%. That is,

$$0.1 = \frac{1}{\sqrt{n}} \sqrt{\text{Var}(\tilde{p}_n)} = \frac{1}{\sqrt{n}} \sqrt{\frac{E[w(Z_1)^2 I\{Z_1 > 3\}]}{p^2} - 1}.$$  

Evaluating the expectation using estimates for tail probabilities in the normal distribution, the number of samples $n$ needed to reach the desired precision is roughly 800. That is, 0.15% of the sample size needed for standard Monte Carlo.

The above is an example of a change of measure, something already encountered in Section 2. In the introduction we mentioned that the idea of importance sampling is to change the dynamics of the system one is trying to model. Since the randomness of the system is described by a probability measure, formally this change in dynamics corresponds to a change of (probability) measure. In the example, the change is from the original distribution $N(0,1)$ to the sampling distribution $N(3,1)$ and the function $w$
used to make the estimator based on the $Z_i$’s unbiased is often referred to as the weight function.

We now give a formal description of importance sampling. Suppose that, for some random variable $X$ with distribution $F$ on a space $\mathcal{X}$, the quantity of interest is the expectation $F(f) = \int f(x) F(dx)$ for some $F$-integrable function $f : \mathcal{X} \rightarrow \mathbb{R}$. Furthermore, let $\tilde{F}$ be the chosen sampling distribution, that is, the new dynamics. It is required that $F \ll \tilde{F}$ on the support of the function $f$. This guarantees the existence of the Radon-Nikodym derivative $dF/d\tilde{F}$ on $\{f > 0\}$, which in turn makes the weight function $w(x) = \frac{dF}{d\tilde{F}}(x)I\{f(x) > 0\}, x \in \mathcal{X}$ well defined. Noting that $F(f) = \int\int f(x) F(dx) = \int\int f(x)\frac{dF}{d\tilde{F}}(x)\tilde{F}(dx) = \tilde{F}(wf)$, it is possible to construct an unbiased estimator of $F(f)$ as

$$\tilde{F}_n(wf) = \frac{1}{n}\sum_{i=1}^{n} f(\tilde{X})w(\tilde{X}),$$

where $(\tilde{X}_1, \ldots, \tilde{X}_n)$ is an independent sample from $\tilde{F}$. The estimator $\tilde{F}_n(wf)$ is the importance sampling estimator of $F(f)$. In the example above, $F$ is $N(0, 1)$, $\tilde{F}$ is $N(3, 1)$ and the weight function $w$ is the ratio between the respective densities. For the two normal distributions, $F \ll \tilde{F}$ holds true on the entire real line (as does $\tilde{F} \ll F$) but in general it suffices that the absolute continuity holds on the set $\{f > 0\}$.

In situations largely determined by rare events standard Monte Carlo can be inefficient. The most obvious example is when estimating probabilities of rare events and this becomes apparent even in the simple toy problem above. Importance sampling is a remedy for this. For this statement to be made precise requires a measure of efficiency. Since importance sampling yields unbiased estimators the canonical choice is variance. Consider first standard Monte Carlo. To make this as illuminating as possible we return to the example of Section 2 concerning a random walk exceeding a threshold. Let us recall the setup. The random variables $X_1, X_2, \ldots$ are real-valued with distribution $F$, $\mathbb{E}[X_1] = 0$ and let $S_n = \sum_{i=1}^{n} X_i$. We are interested in the probability $p_n = \mathbb{P}(S_n \geq na)$ for $a > 0$. Let $F^n$ denote the product measure of $n$ copies of $F$. Staying with the notation from above, the task is to estimate the expectation $F^n(f)$ with $f(x) = I\{\sum_{i=1}^{n} x_i \geq na\}$ for $x \in \mathbb{R}^n$. Take $(X_1^{(i)}, \ldots, X_n^{(i)})$ as independent samples from $F^n$, $i = 1, \ldots, k$, and set $S_n^{(i)} = \sum_{j=1}^{n} X_j^{(i)}$. The Monte Carlo estimator is $\Phi_k^n(f) = (1/k)\sum_{i=1}^{k} I\{S_n^{(i)} \geq na\}$.
which has variance
\[ \text{Var}(F^n_k(f)) = \text{Var}\left(\frac{1}{k} \sum_{i=1}^{k} I\{S^{(i)}_n \geq na\}\right) = \frac{pn(1-pn)}{k}, \]
and relative error
\[ RE_{MC} = \frac{\text{Var}(F^n_k(f))}{pn} = \frac{1}{\sqrt{k}} \sqrt{\frac{1-pn}{pn}}. \]

Ideally, one would want the relative error to be no larger than \( pn \) (in the toy problem above we used 10%) and the number of samples must then be of size \( \approx \frac{1}{p^2 n} \). It is clear that for very small probabilities \( pn \) this could mean an enormous computational burden and put severe stress on the random number generator used. Next, consider the same thing for importance sampling using \( \tilde{F}^n \) as sampling distribution. That is, \( \tilde{S}^{(1)}_n, \ldots, \tilde{S}^{(k)}_n \) are constructed from independent samples from \( \tilde{F}^n \) and the importance sampling estimator of \( pn \) is \( \tilde{F}^n_k(f) = (1/k) \sum_{i=1}^{k} w_n(\tilde{S}^{(i)}_n)I\{\tilde{S}^{(i)}_n \geq na\} \), where \( w_n \) is the weight function (likelihood ratio) between \( F^n \) and \( \tilde{F}^n \). The corresponding relative error is
\[ RE_{IS} = \frac{\text{Var}(\tilde{F}^n_k(f))}{pn} = \frac{1}{\sqrt{k}} \sqrt{\frac{\mathbb{E}[w_n(\tilde{S}^{(i)}_n)^2I\{\tilde{S}^{(i)}_n \geq na\}]}{p^2 n}} - 1. \]

The relative error is determined by the ratio
\[ \frac{\mathbb{E}[w_n(\tilde{S}^{(i)}_n)^2I\{\tilde{S}^{(i)}_n \geq na\}]}{p^2 n}. \]

By Jensen’s inequality this is bounded from below by 1, which would correspond to relative error 0. The lower bound is achieved by the so-called zero-variance change of measure; \( \tilde{F}^n \) is the distribution under \( F^n \) conditioned on the sum exceeding \( na \). Such a change of measure is clearly infeasible in practice as it requires the knowledge of \( pn \), the quantity we are trying to estimate.

In the above setting, a standard optimality criterion is the following. Suppose that for some \( \gamma > 0 \), the sequence of probabilities \( \{pn\} \) satisfies
\[ \lim_{n \to \infty} -\frac{1}{n} \log pn = \gamma. \]

The lower bound \( \mathbb{E}[w_n(\tilde{S}^{(i)}_n)^2I\{\tilde{S}^{(i)}_n \geq na\}] \geq p^2 n \) follows from Jensen’s inequality. This implies that
\[ \lim sup_{n \to \infty} -\frac{1}{n} \log \mathbb{E}[w_n(\tilde{S}^{(i)}_n)^2I\{\tilde{S}^{(i)}_n \geq na\}] \leq 2\gamma. \]
An importance sampling algorithm is called *asymptotically optimal* if the upper bound is achieved, that is if
\[
\liminf_{n \to \infty} -\frac{1}{n} \log \mathbb{E}[w_n(\tilde{S}_n^{(i)})^2 I\{\tilde{S}_n^{(i)} \geq na\}] \geq 2\gamma.
\]
In the literature this is also known as *logarithmically efficient relative error*.

In the above example, if the sampling distribution \( \tilde{F}_n \) is to be the product measure from some \( \tilde{F} \) (that is, independent summands \( \tilde{X}_i \)'s with common distribution \( \tilde{F} \)) the optimal change of measure is the one provided by large deviation analysis. That is, an exponential change of measure such that the random variables all have mean \( a \). This particular choice of sampling algorithm has logarithmically efficient relative error and is in fact the only sampling distribution \( \tilde{F} \) for which that is true ([1], Theorem 2.5). This is what was earlier referred to as the change of measure used to prove the lower bound of the LDP being a good sampling distribution. If one takes into account the interpretation of the rate function (see Section 2) it is intuitively clear that such a choice of sampling distribution should be good - one want to sample in a way such that the event in question becomes as likely as possible. However, recall that this intuition is not “proof” that the algorithm is efficient and indeed there are situations in which it is not.

Even though the ideas used for importance sampling dates back to the 1940’s it is often considered that Siegmund in [27] started the kind of research that is carried out today. In [27] importance sampling is studied in the context of sequential tests and the conclusion is that the change of measure that is used to prove the lower bound of the LDP is the choice for the sampling distribution. The analysis in [27] is related to large deviation theory and the connection between the two topics has since then grown very strong.

Sadowsky unified the work on estimating large deviation probabilities in an abstract setting in [24], the main motivation being sample path problems such as time-varying level crossings. In [19] it is shown that although the change of measure in large deviations often produce sampling distributions with good performance, the method is by no means universal. Indeed, if the conditions of the results established in, e.g., [24] are not met, then the estimators suggested by a large deviation analysis can have really poor performance. Rather than using just one (fixed) change of measure, algorithms can also be state-dependent. In terms of a sum of random variables, the change of measure can evolve dynamically and depend on the previous summands rather than preserving independence. Since efficiency is traditionally expressed in terms of variance it is natural that the design of algorithms has also been done with this in mind. That is, algorithms have been designed in a way so as to, whilst remaining unbiased, control the second moment. For state-dependent algorithms, Dupuis and Wang made important contributions in [13]. They were primarily interested in sums of random variables and observed that asymptotic optimality is related to solutions of an Isaacs
equation. By finding subsolutions to this equation one can design importance sampling algorithms with good performance; see also [14]. Another successful approach to designing state-dependent algorithms is the use of Lyapunov inequalities [3, 4, 5]. However, the emphasis for such algorithms has been on the heavy-tailed setting.

As long as an estimator is unbiased, variance is indeed the canonical measure of efficiency. However, when one wants to estimate quantities in which, for example, the random variables of interest enter in a non-linear way, variance is no longer the obvious choice. This is the case for many applications of stochastic simulation. For example, in finance it is often the tail of a distribution that is of interest. Risk measures such as Value-at Risk (VaR) are used throughout all of finance and insurance. Explicit calculations are not possible and in order to estimate such risk measures stochastic simulation is used. How to design algorithms for such problems? For the specific problem of using importance sampling to compute VaR in finance, [18] investigates a bisection search method. Basically, standard importance sampling algorithms for computing a probability are used to compute the probability of exceeding some high threshold. Depending on how that probability compares to the quantile of interest the threshold is changed and the procedure is run once again. Once the estimated probability coincides with the desired VaR level the algorithm is complete. In [20] Glynn takes a more direct approach and studies central limit theorems for the importance sampling estimators. Efficiency is put in terms of the resulting variance in the limiting normal distribution. More recently, [21] continued on a path similar to that of Glynn and considered weighted empirical processes that are associated with an importance sampling algorithm. By proving central limit theorems for the estimators of VaR and Expected Shortfall in an empirical process setting, the authors are able to relate efficiency for the more complicated problems to the ability of an algorithm to estimate probabilities. In general, the simulation literature is scarce regarding how to quantify efficiency in a good way when the estimators are no longer unbiased or, more generally, when variance is not certain to be obvious choice.

4 Summary of papers

Paper 1: Large deviations for weighted empirical measures arising in importance sampling

In this paper we study large deviation results related to the weighted empirical measures that arise in importance sampling. The aim of the paper is to complement the variance analysis common in efficiency analysis of importance sampling algorithms by studying the rate function associated with the LDP for the weighted empirical measure which can be associated with the output of an importance sampling algorithm.
For the sake of illustration, consider estimating an expectation $F(f)$ for a distribution $F$ and some $F$-integrable function $f$. Using standard Monte Carlo simulation, an estimate is given by

$$F_n(F) = \frac{1}{n} \sum_{i=1}^{n} f(X_i),$$

where the $X_i$’s are independent with common distribution $F$. Suppose now that $\mathbb{E}[\exp\{\theta f(X)\}]$ is finite for $\theta$ in a neighborhood of the origin. Then, Cramér’s theorem ([7], Theorem 2.2.3.) states that $F_n(f)$ satisfies the LDP with rate function $I(x) = \sup_{\theta} \{\theta x - \kappa(\theta)\}$, where

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

Suppose that we want a precision $\varepsilon$ with probability at least $1 - \delta$. That is, $n$ needs to be large enough so that

$$\mathbb{P}(|F_n(f) - F(f)| \geq \varepsilon F(f)) \leq \delta .$$

Cramér’s theorem gives an upper bound in terms of the rate function $I$. If $A_\varepsilon$ is the complement of the open ball of radius $\varepsilon$ around $F(f)$, then at least approximately

$$\mathbb{P}(F_n(f) \in A_\varepsilon^c) \leq e^{-nI(A_\varepsilon)},$$

from which it is concluded that an upper bound $\delta$ on the probability corresponds to

$$n \geq \frac{1}{I(A_\varepsilon)}(-\log \delta).$$

Hence, the computational cost for standard Monte Carlo is expressed in terms of the rate function of Cramér’s theorem. If instead importance sampling, with sampling distribution $\tilde{F}$, is used, an estimator of $F(f)$ is given by

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

where the $\tilde{X}_i$’s are independent with common distribution $\tilde{F}$ and $w = dF/d\tilde{F}$. Let $\tilde{P}$ and $\tilde{E}$ denote the probability and expectation in this setting. Analogous to the case of standard Monte Carlo, if $\tilde{E}[\exp\{\theta w(\tilde{X})f(\tilde{X})\}]$ is finite for $\theta$ in a neighborhood of the origin the Cramér’s theorem states that $\tilde{F}_n(f)$ satisfies the LDP with rate function $I^w(x) = \sup_{\theta} \{\theta x - \kappa^w(\theta)\}$, where

$$\kappa^w(\theta) = \log \tilde{E}[\exp\{\theta w(\tilde{X})f(\tilde{X})\}] .$$
Similar to standard Monte Carlo, in order to reach a precision $\varepsilon$ with probability at least $1 - \delta$ the number of samples needed is roughly

$$n_{IS} \geq \frac{1}{I^w(A_\varepsilon)}(-\log \delta).$$

This gives a way of quantifying the gain in using importance sampling as compared to standard Monte Carlo; compare the rate functions $I$ and $I^w$. Furthermore it provides a possible criteria for choosing the sampling distribution. Namely, one should chose the sampling distribution which maximizes $I^w(A_\varepsilon)$.

To generalize the method here described so as to encompass not only expectations but also more general functionals of a distribution, we suggest studying the empirical measures associated with standard Monte Carlo and importance sampling respectively. For standard Monte Carlo, Sanov’s theorem provides the LDP. However, no such result is available for the weighted empirical measures arising in importance sampling. The main result of the paper, Theorem 3.1, is a Laplace principle for such weighted empirical measures. More precisely, suppose that $f$ is some $F$-integrable function and $w = (dF/d\tilde{F})I\{f > 0\}$. Then, under the assumption that $\tilde{E}[\exp\{\theta w(\tilde{X})f(\tilde{X})\}] < \infty$ for all $\theta > 0$ and an additional technical assumption, the sequence $\{\tilde{F}^w_n\}$ satisfies the Laplace principle on the space of finite measures equipped with the $\tau$-topology.

Aside from establishing the large deviation result for the weighted empirical measures arising in importance sampling, we apply the result to quantify the performance of an importance sampling algorithm over a collection of subsets of a given target set. The idea is that if an algorithm is designed for some target set $A$, then it should also have good performance over subsets $C$ of $A$ that are not given too small probability under the original distribution $F$. This is made precise by studying the rate function of the Laplace principle for the weighted empirical measures over some appropriate subsets in the space of finite measures. The rate function is studied in Lemmas 4.1-4.5 which results in explicit expressions for the quantities of interest, as well as an explicit characterization of the set (in the space of finite measures) that determine performance in this setting. In examples 4.6-4.8 the well-known cases of standard Monte Carlo, (constant) exponential change of measure for a light-tailed random walk and the zero-variance change of measure are analyzed from the large deviation perspective. In these cases the results coincide with that of variance analysis - as they should - and they also provide a measure of the possible gain in computational cost.

The method of proof for the main result is that of the weak convergence approach to large deviations developed by Dupuis and Ellis in [12] and briefly discussed in Section 2.
In this paper we study importance sampling algorithms and the resulting estimators of risk measures from the perspective of empirical processes. Many risk measures can be represented as non-linear functionals of a distribution and therefore the unbiasedness that is often true for Monte Carlo methods does not hold for estimators of such quantities. Therefore, studying solely variance can be misleading regarding the performance of a simulation algorithm. Furthermore, risk measures are often determined mainly by events far out in the tail of the underlying distribution and thus standard Monte Carlo estimates come with a large computational cost, motivating the use of importance sampling.

This paper complements the work presented in [21] on using empirical process results for efficiency analysis of importance sampling algorithms. In [21] the authors study central limit theorems associated with the weighted empirical process arising in importance sampling and we complement that work by studying large deviation properties of the same process. The aim is to establish the theoretical framework necessary for a large deviation approach to efficiency analysis. A nice feature is that, due to the empirical process framework, uniform versions of the LDP is established for estimators of certain quantities.

Let $(E, \mathcal{E})$ be a measurable space and consider some random variables taking their values in this space. Let $\mathcal{F}$ be a collection of functions $f : E \mapsto \mathbb{R}$ and $\ell_\infty(\mathcal{F})$ the space of bounded, real-valued functions on $\mathcal{F}$. Changing notation slightly from the above, suppose that the original distribution of interest is $\mu$ and the sampling distribution is $\nu$. The large deviation properties of standard empirical processes are considered in [32]. There, necessary and sufficient conditions for the LDP are shown for the empirical process $\{b_n(\mu_n - \mu)\}$ as an element in $\ell_\infty(\mathcal{F})$. Here, $\mu_n$ is the empirical measure obtained from $n$ independent samples from $\mu$ and $b_n$ is a particular type of sequence such that $b_n \to \infty$. Let $w = d\mu/d\nu$ which for now is assumed to be well-defined over the entire underlying space. Theorem 3.1 extends the results of [32] to hold also for the weighted empirical process $\{b_n(\nu_n^w - \mu)\}$, where $\nu_n^w$ is the weighted empirical measure given by $\nu_n^w(g) = \int gw\,d\nu_n$ for every bounded, measurable function $g$. The proof uses the results of [32] and [22] and the contraction principle.

With Theorem 3.1 for the weighted empirical processes established we consider estimators of the risk measures VaR and Expected Shortfall. Using a delta method for large deviations established in [17] together with Theorem 3.1, the LDP is derived for importance sampling estimators of the tail of a distribution (Corollary 4.1) as well as of the quantile process (Theorem 4.2). Finally, the LDP for importance sampling estimators of Expected Shortfall is proved in Theorem 4.3 using the result on estimators of the quantile process.
together with the large deviation concept exponentially good approximation. The proofs of the various results utilize both classical large deviation techniques, such as the contraction principle and exponential approximations, and results from the theory of empirical processes and general Banach space valued random variables.

References


