Chapter 11
Importance Sampling for Multi-Constraints Rare Event Probability

Virgile Caron

11.1 Introduction and Context

In this paper, we consider efficient estimation of the probability of large deviations of a multivariate sum of independent, identically distributed, light-tailed, and non-lattice random vectors.

Consider $\mathbf{X}_1^n := (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ n i.i.d. random vectors with known common density $p_{\mathbf{X}}$ on $\mathbb{R}^d$, $d \geqslant 1$, copies of $\mathbf{X} := \left(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(d)}\right)$. The superscript (j) pertains to the coordinate of a vector and the subscript i pertains to replications. Consider also u a measurable function defined from $\mathbb{R}^d$ to $\mathbb{R}^s$. Define $\mathbf{U} := u(\mathbf{X})$ with density $p_{\mathbf{U}}$ and

$$\mathbf{U}_{1,n} := \sum_{i=1}^{n} \mathbf{U}_i .$$

We intend to estimate for large but fixed n

$$P_n := P\left(\mathbf{U}_{1,n} \in nA\right) \tag{11.1}$$

where A is a non-empty measurable set of $\mathbb{R}^s$ such as $E[u(\mathbf{X})] \notin A$. In [3], the authors consider in detail the case where $d = s = 1$, $A := A_n = (a_n, \infty)$ and a_n is a convergent sequence.

The basic estimate of P_n is defined as follows: generate L i.i.d. samples $X_1^n(l)$ with underlying density $p_{\mathbf{X}}$ and define

V. Caron (✉)
Telecom ParisTech, 37-39 rue Dareau, 75014 Paris, France
e-mail: virgile.caron@telecom-paristech.fr; virgile.caron@upmc.fr

V.B. Melas et al. (eds.), *Topics in Statistical Simulation*, Springer Proceedings in Mathematics & Statistics 114, DOI 10.1007/978-1-4939-2104-1_11

$$\widetilde{P_n} := \frac{1}{L}\sum_{l=1}^{L} \mathbb{1}_{\mathscr{E}_n}\left(X_1^n(l)\right)$$

where

$$\mathscr{E}_n := \left\{(x_1,\ldots,x_n) \in \left(\mathbb{R}^d\right)^n : (u(x_1)+\cdots+u(x_n)) \in nA\right\}. \tag{11.2}$$

The Importance Sampling estimator of P_n with sampling density g on $\left(\mathbb{R}^d\right)^n$ is

$$\widehat{P_n} := \frac{1}{L}\sum_{l=1}^{L} \hat{P}_n(l)\mathbb{1}_{\mathscr{E}_n}\left(Y_1^n(l)\right) \tag{11.3}$$

where $\hat{P}_n(l)$ is called "importance factor" and can be written

$$\hat{P}_n(l) := \frac{\prod_{i=1}^{n} p_{\mathbf{X}}(Y_i(l))}{g\left(Y_1^n(l)\right)} \tag{11.4}$$

where the L samples $Y_1^n(l) := (Y_1(l),\ldots,Y_n(l))$ are i.i.d. with common density g; the coordinates of $Y_1^n(l)$ however need not be i.i.d. It is known that the optimal choice for g is the density of $\mathbf{X}_1^n := (\mathbf{X}_1,\ldots,\mathbf{X}_n)$ conditioned upon $\left(\mathbf{X}_1^n \in \mathscr{E}_n\right)$, leading to a zero variance estimator. We refer to [5] for the background of this section.

The state-independent IS scheme for rare event estimation (see [6] or [12]), rests on two basic ingredients: the sampling distribution is fitted to the so-called dominating point (which is the point where the quantity to be estimated is mostly captured; see [11]) of the set to be measured; independent and identically distributed replications under this sampling distribution are performed. More recently, a state-dependent algorithm leading to a strongly efficient estimator is provided by [2] when $d = s$, $u(x) = x$ and A has a smooth boundary and a unique dominating point. Indeed, adaptive tilting defines a sampling density for the $i-$th r.v. in the run which depends both on the target event $(\mathbf{U}_{1,n} \in nA)$ and on the current state of the path up to step $i-1$. Jointly with an ad hoc stopping rule controlling the excursion of the current state of the path, this algorithm provides an estimate of P_n with a coefficient of variation independent upon n. This result shows that nearly optimal estimators can be obtained without approximating the conditional density.

The main issue of the method described above is to find dominating point. However, when the dimension of the set A increases, finding a dominating point can be very tricky or even impossible. A solution will be to divide the set under consideration into smaller subset and, for each one of this subset, find a dominating point. Doing so makes the implementation of an IS scheme harder and harder as the dimension increases.

Our proposal is somehow different since it is based on a sharp approximation result of the conditional density of long runs. The approximation holds for any point conditioning of the form $(\mathbf{U}_{1,n} = nv)$. Then sampling v in A according to the distribution of $\mathbf{U}_{1,n}$ conditioned upon $(\mathbf{U}_{1,n} \in nA)$ produces the estimator. By its very definition this procedure does not make use of any dominating point, since it randomly explores the set A. Indeed, our proposal hints on two choices: first do not make use of the notion of dominating point and explore all the target set instead (no part of the set A is neglected); secondly, do not use i.i.d. replications, but merely sample long runs of variables under a proxy of the optimal sampling scheme.

We will propose an IS sampling density which approximates this conditional density very sharply on its first components $y_1, \ldots, y_k$ where $k = k_n$ is very large, namely $k/n \to 1$. However, but in the Gaussian case, k should satisfy $(n-k) \to \infty$ by the very construction of the approximation. The IS density on $\left(\mathbb{R}^d\right)^n$ is obtained multiplying this proxy by a product of a much simpler state-independent IS scheme following [13].

The paper is organized as follows. Section 11.2 is devoted to notations and hypothesis. In Sect. 11.3, we expose the approximation scheme for the conditional density of $\mathbf{X}_1^k$ under $(\mathbf{U}_{1,n} = nv)$. Our IS scheme is introduced in Sect. 11.4. Simulated results are presented in Sect. 11.5 which enlighten the gain of the present approach over state-dependent Importance Sampling schemes.

We rely on [7] where the basic approximation (and proofs) used in the present paper can be found. The real case is studied in [4] and applications for IS estimators can be found in [3].

11.2 Notations and Hypotheses

We consider approximations of the density of the vector $\mathbf{X}_1^k$ on $\left(\mathbb{R}^d\right)^k$, when the conditioning event writes (11.1) and $k := k_n$ is such that

$$0 \leqslant \limsup_{n\to\infty} \frac{k}{n} \leqslant 1 \tag{K1}$$

$$\lim_{n\to\infty} (n-k) = +\infty. \tag{K2}$$

Therefore we may consider the asymptotic behavior of the density of the random walk on long runs.

Throughout the paper the value of a density $p_{\mathbf{Z}}$ of some continuous random vector $\mathbf{Z}$ at point z may be written $p_{\mathbf{Z}}(z)$ or $p\,(\mathbf{Z} = z)$, which may prove more convenient according to the context.

Let p_{nv} (and distribution P_{nv}) denote the density of $\mathbf{X}_1^k$ under the local condition $(\mathbf{U}_{1,n} = nv)$

$$p_{nv}\left(\mathbf{X}_1^k = Y_1^k\right) := p(\mathbf{X}_1^k = Y_1^k \,|\, \mathbf{U}_{1,n} = nv) \tag{11.5}$$

where Y_1^k belongs to $\left(\mathbb{R}^d\right)^k$ and v belongs to A.

We will also consider the density p_{nA} (and distribution P_{nA}) of $\mathbf{X}_1^k$ conditioned upon $(\mathbf{U}_{1,n} \in nA)$

$$p_{nA}\left(\mathbf{X}_1^k = Y_1^k\right) := p(\mathbf{X}_1^k = Y_1^k \,|\, \mathbf{U}_{1,n} \in nA). \tag{11.6}$$

The approximating density of p_{nv} is denoted g_{nv}; the corresponding approximation of p_{nA} is denoted g_{nA}. Explicit formulas for those densities are presented in the next section.

11.3 Multivariate Random Walk Under a Local Conditioning Event

Let ε_n be a positive sequence such as

$$\lim_{n\to\infty} \varepsilon_n^2(n-k) = \infty \tag{E1}$$

$$\lim_{n\to\infty} \varepsilon_n(\log n)^2 = 0 \tag{E2}$$

It will be shown that $\varepsilon_n\,(\log n)^2$ is the rate of accuracy of the approximating scheme.

We assume that $\mathbf{U} := u(\mathbf{X})$ has a density $p_{\mathbf{U}}$ (with p.m. $P_{\mathbf{U}}$) absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^s$. Furthermore, we assume that u is such that the characteristic function of $\mathbf{U}$ belongs to L^r for some $r \geqslant 1$.

Denote $\underline{0}$ is the vector of $\mathbb{R}^s$ with all coordinates equal to 0 and $V(\underline{0})$ a neighborhood of $\underline{0}$.

We assume that $\mathbf{U}$ satisfy the Cramer condition, meaning

$$\Phi_{\mathbf{U}}(t) := E[\exp < t, \mathbf{U} >] < \infty, \;\; t \in V(\underline{0}) \subset \mathbb{R}^s.$$

and define

$$m(t) := {}^t\nabla \log(\Phi_{\mathbf{U}}(t)), \;\; t \in V(\underline{0}) \subset \mathbb{R}^s$$

and

$$\varkappa(t) := {}^t\nabla\nabla \log(\Phi_{\mathbf{U}}(t)), \;\; t \in V(\underline{0}) \subset \mathbb{R}^s.$$

as the mean and the covariance matrix of the tilted density defined by

$$\pi_u^{\alpha}(x) := \frac{\exp < t, u(x) >}{\Phi_{\mathbf{U}}(t)} p_{\mathbf{X}}(x). \tag{11.7}$$

where t is the only solution of $m(t) = \alpha$ for α in the convex hull of $P_{\mathbf{U}}$. Conditions on $\Phi_{\mathbf{U}}(t)$ which ensure existence and uniqueness of t are referred to steepness properties (see [1], p153 ff, for all properties of moment generating function used in this paper).

We now state the general form of the approximating density. Let $v \in A$ and denote

$$g_0(y_1|y_0) := \pi_u^{v}(y_1) \tag{11.8}$$

with an arbitrary y_0 and π_u^v defined in (11.7).

For $1 \leqslant i \leqslant k-1$, we recursively define $g(y_{i+1}|y_1^i)$. Set $t_i \in \mathbb{R}^s$ to be the unique solution to the equation

$$m(t_i) = m_{i,n} := \frac{n}{n-i}\left(v - \frac{u_{1,i}}{n}\right) \tag{11.9}$$

where $u_{1,i} = u(y_1) + \cdots + u(y_i)$.

Denote

$$\varkappa_{(i,n)}^{j,l} := \frac{d^2}{dt^{(j)}dt^{(l)}}\left(\log E_{\pi_{\mathbf{U}}^{m_{i,n}}} \exp < t, \mathbf{U} >\right)(\underline{0})$$

and

$$\varkappa_{(i,n)}^{j,l,m} := \frac{d^3}{dt^{(j)}dt^{(l)}dt^{(m)}}\left(\log E_{\pi_{\mathbf{U}}^{m_{i,n}}} \exp < t, \mathbf{U} >\right)(\underline{0}).$$

for j, l and m in $\{1, \ldots, s\}$. In the sequel, $\varkappa_{(i,n)}$ will denote the matrix with elements $\left(\varkappa_{(i,n)}^{j,l}\right)_{1\leqslant j,l\leqslant s}$.

Denote

$$g(y_{i+1}|y_1^i) := C_i \mathfrak{n}_s\left(u(y_{i+1}); \beta\alpha + v, \beta\right) p_{\mathbf{X}}(y_{i+1}) \tag{11.10}$$

where C_i is a normalizing factor, $\mathfrak{n}_s\left(u(y_{i+1}); \beta\alpha + v, \beta\right)$ is the normal density at $u(y_{i+1})$ with mean $\beta\alpha + v$ and covariance matrix β. α and β are defined by

$$\alpha := \left(t_i + \frac{\varkappa_{(i,n)}^{-2}\gamma}{2(n-i-1)}\right)$$

and

$$\beta := \varkappa_{(i,n)}(n-i-1)$$

and γ defined by

$$\gamma := \left(\sum_{j=1}^{s} \varkappa_{(i,n)}^{j,j,p} \right)_{1 \leqslant p \leqslant s}.$$

Then

$$g_{nv}(y_1^k) := g_0(y_1|y_0) \prod_{i=1}^{k-1} g(y_{i+1}|y_1^i) \tag{11.11}$$

Theorem 1. *Assume (E1), (E2), (K1) and (K2).*

- *Let Y_1^k be a sample from density p_{nv}. Then*

$$p\left(\boldsymbol{X}_1^k = Y_1^k | \mathbf{U}_{1,n} = nv\right) = g_{nv}(Y_1^k)(1 + o_{P_{nv}}(1 + \varepsilon_n (\log n)^2)) \tag{11.12}$$

- *Let Y_1^k be a sample from density g_{nv}. Then*

$$p\left(\boldsymbol{X}_1^k = Y_1^k | \mathbf{U}_{1,n} = nv\right) = g_{nv}(Y_1^k)(1 + o_{G_{nv}}(1 + \varepsilon_n (\log n)^2)) \tag{11.13}$$

Remark 11.1. The approximation of the density of $\mathbf{X}_1^k$ is not performed on the sequence of entire spaces $\left(\mathbb{R}^d\right)^k$ but merely on a sequence of subsets of $\left(\mathbb{R}^d\right)^k$ which contains the trajectories of the conditioned random walk with probability going to 1 as n tends to infinity. The approximation is performed on *typical paths*. For the sake of applications in Importance Sampling, (11.13) is exactly what we need. Nevertheless, as proved in [7], the extension of our results from typical paths to the whole space $\left(\mathbb{R}^d\right)^k$ holds: convergence of the relative error on large sets imply that the total variation distance between the conditioned measure and its approximation goes to 0 on the entire space.

Remark 11.2. The rule which defines the value of k for a given accuracy of the approximation is stated in Sect. 5 of [7].

Remark 11.3. When the $\mathbf{X}_i$'s are i.i.d. multivariate Gaussian with diagonal covariance matrix and $u(x) = x$, the results of the approximation theorem are true for $k = n - 1$ without the error term. Indeed, it holds $p(\mathbf{X}_1^{n-1} = x_1^{n-1} \big| \mathbf{U}_{1,n} = nv) = g_{nv}\left(x_1^{n-1}\right)$ for all x_1^{n-1} in $\left(\mathbb{R}^d\right)^{n-1}$.

As stated above the optimal choice for the sampling density is p_{nA}. It holds

$$p_{nA}(x_1^k) = \int_A p_{nv}\left(\mathbf{X}_1^k = x_1^k\right) p(\mathbf{U}_{1,n}/n = v | \, \mathbf{U}_{1,n} \in nA) dv \tag{11.14}$$

so that, in contrast with [2] or [6], we do not consider the dominating point approach but merely realize a sharp approximation of the integrand at any point of A and consider the dominating contribution of all those distributions in the evaluation of the conditional density p_{nA}.

11.4 Adaptive IS Estimator for Rare Event Probability

The IS scheme produces samples $Y := (Y_1, \ldots, Y_k)$ distributed under g_{nA}, which is a continuous mixture of densities g_{nv} as in (11.11) with $p\left(\mathbf{U}_{1,n}/n = v|\mathbf{U}_{1,n} \in nA\right)$.

Simulation of samples $\mathbf{U}_{1,n}/n$ under this density can be performed through Metropolis–Hastings algorithm, since

$$r(v, v') := \frac{p(\mathbf{U}_{1,n}/n = v|\,\mathbf{U}_{1,n} \in nA)}{p(\mathbf{U}_{1,n}/n = v'|\,\mathbf{U}_{1,n} \in nA)}$$

turns out to be independent upon $P\left(\mathbf{U}_{1,n} \in nA\right)$. The proposal distribution of the algorithm should be supported by A.

The density g_{nA} is extended from $\left(\mathbb{R}^d\right)^k$ onto $\left(\mathbb{R}^d\right)^n$ completing the $n-k$ remaining coordinates with i.i.d. copies of r.v's $Y_{k+1}, \ldots, Y_n$ with common tilted density

$$g_{nA}\left(y_{k+1}^n\middle|\, y_1^k\right) := \prod_{i=k+1}^{n} \pi_u^{m_k}(y_i) \tag{11.15}$$

with $m_k := m(t_k) = \frac{n}{n-k}\left(v - \frac{u_{1,k}}{n}\right)$ and

$$u_{1,k} = \sum_{i=1}^{k} u(y_i).$$

The last $n-k$ r.v's $\mathbf{Y}_i$'s are therefore drawn according to the state independent i.i.d. scheme in phase with Sadowsky and Bucklew [13].

We now define our IS estimator of P_n. Let $Y_1^n(l) := Y_1(l), \ldots, Y_n(l)$ be generated under g_{nA}. Let

$$\widehat{P_n}(l) := \frac{\prod_{i=1}^n p_{\mathbf{X}}(Y_i(l))}{g_{nA}(Y_1^n(l))} \mathbb{1}_{\mathcal{E}_n}\left(Y_1^n(l)\right) \tag{11.16}$$

and define

$$\widehat{P_n} := \frac{1}{L} \sum_{l=1}^{L} \widehat{P_n}(l). \tag{11.17}$$

in accordance with (11.3).

Remark 11.4. In the real case and for $A = (a, \infty)$, the authors of [3] show that under certain regularity conditions the resulting relative error of the estimator is proportional to $\sqrt{n-k_n}$ and drops by a factor $\sqrt{n-k_n}/\sqrt{n}$ with respect to the state independent IS scheme. In [8], the authors propose a slight modification in the extension of g_{nA} which allows to prove the strong efficiency of the estimator (11.17) using arguments from both [2] and [3].

11.5 When the Dimension Becomes Very High

This section compares the performance of the present approach with respect to the standard tilted one using i.i.d. replications under (11.7) on an extension of a well-known example developed in [9] and in [10]. Let $B := (\mathscr{E}_{100})^d$ which is the d-Cartesian product of $\mathscr{E}_{100}$ defined by

$$\mathscr{E}_{100} := \left\{ x_1^{100} : \frac{|x_1 + \cdots + x_{100}|}{100} > 0.28 \right\}.$$

We want to estimate $P_{100} = P[B]$ and explore the gain in relative accuracy when the dimension of the measured set increases. Consider 100 r.v.'s X_i 's i.i.d. random vectors in $\mathbb{R}^d$ with common i.i.d. $N(0.05, 1)$ distribution. Our interest is to show that in this simple asymmetric case our proposal provides a good estimate, while the standard IS scheme ignores a part of the event B. The standard i.i.d. IS scheme introduces the dominating point $a =^t (0.28, \ldots, 0.28)$ and the family of i.i.d. tilted r.v's with common $N(a, 1)$ distribution. It can be seen that a large part of B is never visited through the procedure, inducing a bias in the estimation. Indeed, the *rogue path curse* (see [9]) produces an overwhelming loss in accuracy, imposing a very large increase in runtime to get reasonable results. Under the present proposal the distribution of the Importance Factor concentrates around P_{100} avoiding *rogue path*.

This example is not as artificial as it may seem; indeed, it leads to a 2^d dominating points situation which is quite often met in real life. Exploring at random the set of interest avoids any search for dominating points. Drawing L i.i.d. points $v_1, \ldots, v_L$ according to the distribution of $\mathbf{U}_{1,100}/100$ conditionally upon B we evaluate P_{100} with $k = 99$; note that in the Gaussian case Theorem 1 provides an exact description of the conditional density of X_1^k for all k between 1 and n. The following figure shows the gain in relative accuracy w.r.t. the state independent IS scheme according to the growth of d. The value of P_{100} is 10^{-2d} (Fig. 11.1).

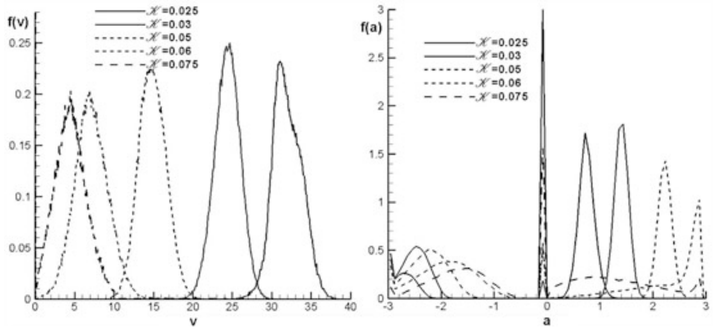

Fig. 11.1 Relative Accuracy of the adaptive estimate (*dotted line*) w.r.t. i.i.d. tilted one (*solid line*) as a function of the dimension d for $L = 1{,}000$

Conclusion

In this paper, we explore a new way to estimate multi-constraints large deviation probability. In future work, the author will investigate the theoretical behavior of the relative error of our proposed estimator.

References

1. Barndorff-Nielsen, O.E.: Information and Exponential Families in Statistical Theory. Wiley, New York (1978)
2. Blanchet, J.H., Glynn, P.W., Leder, K.: Efficient simulation of light-tailed sums: an old-folk song sung to a faster new tune.... In: L'Ecuyer, P., Owen, A.B. (eds.) Monte Carlo and Quasi-Monte Carlo Methods, pp. 227–248. Springer, Berlin (2009)
3. Broniatowski, M., Caron, V.: Towards zero variance estimators for rare event probabilities. ACM TOMACS: Special Issue on Monte Carlo Methods in Statistics **23**(1) (2013) [article 7]
4. Broniatowski, M., Caron, V.: Long runs under a conditional limit distribution. Ann. Appl. Probab (2014, to appear)
5. Bucklew, J.A.: Introduction to Rare Event Simulation. Springer Series in Statistics. Springer, New York (2004)
6. Bucklew, J.A., Ney, P., Sadowsky, J.S.: Monte Carlo simulation and large deviations theory for uniformly recurrent markov chains. J. Appl. Probab. **27**(11), 49–61 (1990)
7. Caron, V.: Approximation of a multivariate conditional density (2013). arxiv:1401.3256
8. Caron, V., Guyader, A., Munoz, Z.M., Tuffin, B.: Some recent results in rare event estimation. In: ESAIM Proceedings (2013, to appear)
9. Dupuis, P., Wang, H.: Importance sampling, large deviations, and differential games. Stoch. Stoch. Rep. **76**, 481–508 (2004)
10. Glasserman, P., Wang, Y.: Counterexamples in importance sampling for large deviations probabilities. Ann. Appl. Probab. **7**(3), 731–746 (1997)
11. Ney, P.: Dominating points and the asymptotics of large deviations for random walk on $\mathbb{R}^d$. Ann. Probab. **11**(1), 158–167 (1983)

12. Sadowsky, J.S.: On Monte-Carlo estimation of large deviations probabilities. Ann. Appl. Probab. **9**(2), 493–503 (1996)
13. Sadowsky, J.S., Bucklew, J.A.: On large deviations theory and asymptotically efficient Monte Carlo estimation. IEEE Trans. Inform. Theory **36**(3), 579–588 (1990)