ROYAL INSTITUTE OF TECHNOLOGY

Large deviations for the weighted empirical measures of importance sampling

Pierre Nyquist & Henrik Hult

Royal Institute of Technology, Stockholm, Sweden

Background

This work is motivated by the task of quantifying the efficiency of importance sampling algorithms. We propose that when importance sampling is used for estimating general functionals of a distribution, not just expectations, efficiency can be expressed in terms of the rate function associated with large deviations of the underlying empirical measures.

Consider estimating the probability $p_a := P(A)$ using CMC, where A is some rare event for the underlying distribution F. An example is $p_a = \mathbf{P}(X_1 + ... + X_n > an)$ for some high threshold a. The CMC estimate $\hat{p}_a$ based on N simulations has relative error

$$RE(\hat{p}_a) = \frac{1}{\sqrt{N} p_a} \sqrt{p_a(1-p_a)},$$

and when p_a is small one would need $N \approx \frac{1}{p_a}$ in order for $RE(\hat{p}_a)$ to be of size comparable to p_a, thus rendering the CMC an inefficient method for rare events. A popular variance reduction method is importance sampling (IS). The idea is to switch to a sampling distribution $\tilde{F}$ under which the event A becomes less rare. The efficiency of this method is determined by the choice of $\tilde{F}$ and two commonly used criteria are *bounded relative error*,

$$\limsup_{a \to \infty} \frac{Var(X)}{p_a^2} < \infty,$$

and *asymptotically optimal relative error*,

$$\limsup_{a \to \infty} \frac{\mathbf{E}[\hat{p}_a^2]}{p_a^2} = 1.$$

Here X comes from the chosen sampling distribution $\tilde{F}$ and $\hat{p}_a$ is the IS estimate of p_a. These criteria stem from the fact that the IS estimator is unbiased and therefore it suffices, and is convenient, to consider the variance.

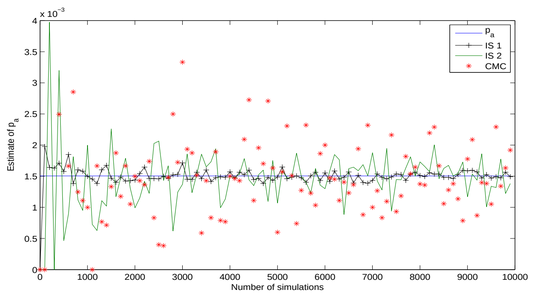

Figure 1: A comparison of the performance of CMC and two different IS algorithms for estimating $p_a = \mathbf{P}(X_1 + ... + X_n > an)$. Here the X_i's are i.i.d. exponentially distributed with mean 1, $n = 10$ and $a = 2.2$. Notice the difference in performance of IS for different choices of $\tilde{F}$.

Problem - quantifying efficiency

The idea behind Monte Carlo estimation of $\Phi(F)$ is to construct the empirical measure $\mathbf{F}_n = \frac{1}{n}\sum_i \delta_{X_i}$ from an i.i.d. sample $X_1, ..., X_n$ from F and take the plug-in estimate $\Phi(\mathbf{F}_n)$. Here δ_x denotes a point mass at x. If $\Phi(F)$ is to a large extent determined by rare events (with respect to F) CMC might again be an inefficient method in terms of computational cost.

When applying IS to estimate $\Phi(F)$ one is in need of a criteria of efficiency in order to make a good choice of sampling distribution $\tilde{F}$. Depending on the functional Φ it might be that expressing efficiency in terms of the variance of the estimator no longer suffices or it is no longer a straightforward task to compute this variance. Therefore alternative notions of efficiency might be needed.

The idea behind this work is to look at the random (weighted) measures that are associated with a simulation algorithm in order to determine the performance.

The weighted empirical measures of IS

Suppose that IS using a sampling distribution $\tilde{F}$ is proposed for estimating $\Phi(F)$. For this to be tractable, it must hold that $F \ll \tilde{F}$ on the part of the underlying space $\mathcal{X}$ that is of interest. We introduce a so called *importance function* $f : \mathcal{X} \to \mathbf{R}$ that characterizes the importance of different regions of $\mathcal{X}$ for evaluating $\Phi(F)$. Assume that $F \ll \tilde{F}$ on the support of f and denote by w the Radon-Nikodym derivative $dF/d\tilde{F}$ on this set. Just as for CMC there is an empirical measure corresponding to IS, namely

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

where $X_1, ..., X_n$ is an i.i.d. sample from $\tilde{F}$. The idea is that if $\tilde{\mathbf{F}}_n^{wf}$ is close to the measure F^f, defined by

$$F^f(g) = \int_{\mathcal{X}} g(x) f(x) F(dx),$$

for each bounded, measurable function g, then $\Phi(\tilde{\mathbf{F}}_n^{wf})$ should be a good estimate of $\Phi(F)$. Through large deviation results for $\{\tilde{\mathbf{F}}_n^{wf}\}$ the efficiency of the algorithm can be expressed in terms of the corresponding rate function.

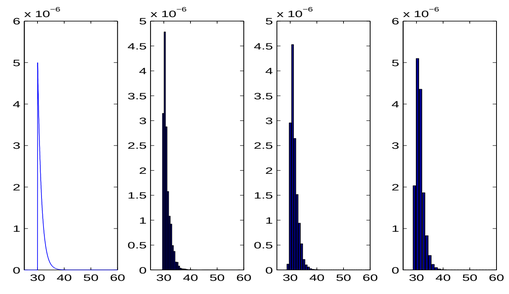

Figure 2: Illustration of the convergence of the weighted empirical measures to F^f, $n = 100, 1000, 10000$. The leftmost function is the density corresponding to F^f. The setup is the same as in Figure 1.

Main result

Denote by $\mathcal{M}$ the space of all finite measures on $\mathcal{X}$ and by $\mathcal{M}_1$ the space of all subprobability measures on $\mathcal{X}$. The main theoretical result of our work is to establish the Laplace principle for $\tilde{\mathbf{F}}_n^{wf}$.

Theorem: *Let F and $\tilde{F}$ be given as above and let f be the importance function chosen for the functional Φ. Under the assumption that $\int e^{\sigma w(x) f(x)} d\tilde{F}(x) < \infty$ for all $\sigma > 0$ and some additional technical conditions, the sequence $\{\tilde{\mathbf{F}}_n^{wf}\}$ of weighted empirical measures satisfies the following Laplace principle on $\mathcal{M}$ equipped with the τ-topology: For all bounded, continuous functionals $h : \mathcal{M} \to \mathbf{R}$,*

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbf{E}[e^{-nh(\tilde{\mathbf{F}}_n^{wf})}] = - \inf_{\nu \in \mathcal{M}} \{I(\nu) + h(\nu)\}.$$

The rate function $I : \mathcal{M} \to [0, \infty]$ is given by

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

where $\Gamma = \{Q \in \mathcal{M}_1 : \mathcal{H}(Q \mid \tilde{F}) < \infty, Q(wf) < \infty\}$ and the mapping Ψ maps a probability measure G, such that $G(wf) < \infty$, to the finite measure G^{wf}.

Application

One potential way in which the result can be applied is as follows. Let $A_\epsilon \subset \mathcal{M}$ be a set that relates to the accuracy of the IS estimator. For example, this could be that the relative error of the estimate is $> \epsilon$. From the derived Laplace principle it follows that for large n,

$$\mathbf{P}(\tilde{\mathbf{F}}_n^{wf} \in A_\epsilon) \approx e^{-nI(A_\epsilon)},$$

and from this the number of samples needed to have a bound δ on the probability of being in the set A_ϵ is roughly

$$n_{IS} = \frac{1}{I(A_\epsilon)}(-\log \delta).$$

In the case that the notion of a set A_ϵ is too restrictive, the functional h of the Laplace principle can be used to make the notion of an efficient algorithm "more smooth". Moreover, we have the following result:

Proposition: *For nice choices of the functional h, the optimization problem in the RHS of the Laplace principle can be expressed as a minimization problem over a real-valued parameter λ.*

References

- Asmussen, S. and Glynn, P. W. - *Stochastic simulation: Algorithms and analysis* (2007)
- Dembo, A. and Zeitouni, O. - *Large deviations techniques and applications*, 2nd ed. (1998)
- Dupuis, P. and Ellis, R. S. - *A weak convergence approach to the theory of large deviations* (1997)

Contact

Pierre Nyquist, *pierren@kth.se*
Henrik Hult, *hult@kth.se*
Department of Mathematics,
Royal Institute of Technology (KTH),
Stockholm, Sweden