Large deviations for weighted empirical measures from importance sampling

Pierre Nyquist

Department of Mathematics
Royal Institute of Technology, Stockholm

RESIM 2012
June 27, 2012

Joint work with Henrik Hult

Introduction

How can large deviation results on the empirical measure level be useful for theoretical analysis of efficiency of importance sampling algorithms?

Introduction

Problem setting

Let X be a random variable, distribution F, taking values in some space $\mathcal{X}$. Consider the task of computing $\Phi(F)$ for some functional Φ:

- Expectation: $\Phi_h(F) = \int hdF =: F(h)$, for some $h : \mathcal{X} \mapsto \mathcal{R}$,
- Quantile: $\Phi_q(F) = F^{-1}(q) = \inf\{x : F((x, \infty)) \leq q\}$, $q \in (0, 1)$,
- L-statistic: $\Phi(F) = \int_0^1 \phi(q)F^{-1}(q)dq$.

When explicit computation is impossible, turn to simulation.

Introduction

Standard Monte Carlo and importance sampling

- Standard Monte Carlo: Sample $X_1, \ldots, X_n$ i.i.d. from F and construct the empirical measure

$$\mathbb{F}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

- Importance sampling: Sample $X_1, \ldots, X_n$ i.i.d. from the *sampling distribution* G, $F \ll G$. Construct the weighted empirical measure,

$$\mathbb{G}_n^w = \frac{1}{n} \sum_{i=1}^{n} w(X_i)\delta_{X_i}, \;\; w := \frac{dF}{dG}.$$

- Corresponding estimates are $\Phi(\mathbb{F}_n)$ and $\Phi(\mathbb{G}_n^w)$

Performance analysis

Introduction

- Performance of importance sampling determined by the choice of sampling distribution G. Typically evaluated in terms of $\mathrm{Var}(\Phi(\mathbb{G}_n^w))$
- Estimation of large deviation probabilities a well studied problem. Large deviation results have been used extensively for studying the rate of the decay of the variance and designing efficient algorithms

Performance analysis

Main idea

- Rather than sample path large deviations, use large deviation results for the empirical measures.
- Replace variance by the rate function of a large deviation principle as the number of particles increases (large sample limit).

Performance of Monte Carlo

Estimation of an expectation

- Let $A_\epsilon(F(h)) = \{x : |x - F(h)| > \epsilon F(h)\}$. For $\Phi(\mathbb{F}_n) = \mathbb{F}_n(h)$, Cramér's theorem provides, for n sufficiently large, the approximation

$$\mathbf{P}(\mathbb{F}_n(h) \in A_\epsilon(F(h))) \approx \exp\{-n \inf_{x \in A_\epsilon(F(h))} \Lambda^*(x)\}.$$

- Number of particles roughly needed for an upper bound δ on the probability:

$$n \approx \frac{1}{\inf_{x \in A_\epsilon(F(h))} \Lambda^*(x)}(-\log \delta).$$

Performance of Monte Carlo

Estimation of a general functional

- For general Φ the idea is to consider the probability of $\mathbb{F}_n$ being close to F. Sanov's theorem provides an LDP for the empirical measures $\mathbb{F}_n$ of Monte Carlo.
- $A_\epsilon \subset \mathcal{M}_1$ a set that relates to the accuracy of $\Phi(\mathbb{F}_n)$. By Sanov's theorem,

$$\mathbf{P}(\mathbb{F}_n \in A_\epsilon) \approx \exp\{-n \inf_{G \in \overline{A}_\epsilon} \mathcal{H}(G \mid F)\}.$$

- Example: $A_\epsilon = \{G \in \mathcal{M}_1 :\mid \Phi(G) - \Phi(F) \mid > \epsilon \Phi(F)\}$.

Performance of Monte Carlo

Estimation of a probability

- Let h be the indicator of some set A, with $p := F(A)$.
- With A_ϵ the ball of radius ϵp centered at p,

$$\inf_{G \in A_\epsilon} \mathcal{H}(G|F) \sim \alpha(\epsilon) p \quad \text{as } p \to 0.$$

- Consistent with analysis of variance of estimator.

Performance of importance sampling

A possible approach

- Suppose $\mathbb{G}_n^w$ satisfies an LDP. Let $A_\epsilon \subset \mathcal{M}$ be some set that relates to the accuracy of the estimate $\Phi(\mathbb{G}_n^w)$.
- The LDP implies, for sufficiently large n,

$$\mathbf{P}(\mathbb{G}_n^w \in A_\epsilon) \approx \exp\{-n \inf_{\nu \in \overline{A}_\epsilon} I^w(\nu)\}.$$

- With δ the desired upper bound for the probability,

$$n \approx \frac{1}{\inf_{\nu \in \overline{A}_\epsilon} I^w(\nu)}(-\log \delta).$$

Performance of importance sampling

A possible approach

- Similarly, the probability of $\mathbb{G}_n^w$ having some undesireable shape $\nu \in \mathcal{M}$ can be studied using $I^w(\nu)$.
- For $\mathbb{G}_n^w(h)$ Cramér's theorem is applicable.
- Sanov's theorem not applicable for the weighted empirical measures $\mathbb{G}_n^w$.
- Need an LDP for $\mathbb{G}_n^w$ in order to quantify the notion of the weighted empirical measures being close to F.

Large deviations for importance sampling

Framework

- Suffices to have the weighted empirical measures $\mathbb{G}_n^w$ close to F in the region that largely determines $\Phi(F)$.
- Let f be an F-integrable function characterizing the importance of different regions of the space $\mathcal{X}$ - *importance function*. Want

$$\mathbb{G}_n^{wf} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)f(X_i)\delta_{X_i},$$

to be close to F^f, where F^f is defined as

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

for each bounded, measurable function g.

Large deviations for importance sampling

Laplace principle

- $\Gamma = \{Q \in \mathcal{M}_1 : \mathcal{H}(Q \mid G) < \infty, Q(wf) < \infty\}$.
- $\Psi : \Gamma \mapsto \mathcal{M}$ the mapping $G \mapsto G^{wf}$.
- $I(\nu) = \inf\{\mathcal{H}(Q \mid G) : \Psi(Q) = \nu, Q \in \Gamma\}$.

Theorem Let F, G and f be given as above, with $F \ll G$ on the support of f. Suppose that $\int e^{wf} dG, \int e^{w^2 f^2} dG < \infty$. Then, for any bounded, continuous $h : \mathcal{M} \mapsto \mathcal{R}$,

$$\lim_n \frac{1}{n} \log \mathbb{E}[e^{-nh(\mathbb{G}_n^{wf})}] = -\inf_{\nu \in \mathcal{M}} \{h(\nu) + I(\nu)\}.$$

Performance of importance sampling

Performance criteria

- The choice of A_ϵ, or ν, reflects your criteria for good performance.
- Suppose that $A \subset \mathcal{X}$ is the region of interest, $\delta \propto F(A)$. A possible choice is

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

Performance of importance sampling

Estimation of a probability

- Let $X_1, X_2, \ldots$ be i.i.d. F and consider the random walk $S_n = \sum_{i=1}^n X_i$. Want to estimate $p_n = \mathbf{P}(S_n/n \geq a)$ for $\mathbb{E}[X_1] < a$.
- Use constant exponential tilting, parameter θ, and the previous choice of A_ϵ; $A = \{S_n/n \geq a\}$.
- Value of the rate function related to $\mathbb{E}[w^{-1}(X)1\{X \in A\}]$. Obtain a lower bound which is maximized for θ such that

$$\kappa'(\theta) = a.$$

Performance of Monte Carlo methods

Rare event limit

- For both Monte Carlo and importance sampling, study the rare event limit,

$$\liminf_{p \to 0} \inf_{\nu \in A_\epsilon} I(\nu).$$

- If the above is zero, use the asymptotic rate $\gamma(p)$ as a measure of efficiency:

$$\inf_{\nu \in A_\epsilon} I(\nu) \sim \gamma(p) \text{ as } p \to 0.$$

Applications for performance analysis

Possible ways to use the large deviation result:

- Comparison of Monte Carlo and importance sampling in terms of the rate functions of large deviation principles.
- Larger rate suggests improved performance.
- The large deviation heuristics of the most likely way for an event to occur can possibly help in designing algorithms that meet the specified performance criteria

Summary

- Propose a way to use the rate function of large deviation results to quantify the performance of importance sampling algorithms.
- Derive a Laplace principle for the weighted empirical measures of importance sampling as the number of particles increases.

Large deviations for importance sampling

Idea of proof

- Relies on the weak convergence approach to large deviations[1].
- Identify $W_n = -\frac{1}{n}\log \mathbb{E}[\exp\{(-nh(\mathbb{G}_n^{wf})\}]$ as the total cost of a stochastic control problem and derive a representation formula.
- The Laplace principle upper bound

$$\limsup_n \frac{1}{n}\log \mathbb{E}[e^{-nh(\mathbb{G}_n^{wf})}] \leq -\inf_{\nu\in\mathcal{M}}\{h(\nu)+I(\nu)\},$$

requires the most work compared to the case of ordinary empirical measures (Sanov's theorem).

[1] Dupuis and Ellis (1997)

Performance of importance sampling

Estimation of a probability, cont'd.

- With the described choice of A_ϵ, minimizing the rate I corresponds to minimizing $G(C)$ for $C \subset A$, $F(C) \geq \delta$ such that the condition on the Radon-Nikodym derivative is fulfilled.
- Possible to explicitly characterize the optimal $\tilde{C}$ in terms of the weight function.
- Optimal rate is the obtained by finding the G, within some prescribed family, which maximizes

$$G(\tilde{C}) = \mathbb{E}[w^{-1}(X)1\{X \in A\}].$$