Semi-Parametric Importance Sampling for Rare-event probability Estimation

Z. I. Botev and P. L'Ecuyer

IMACS Seminar 2011

Borovets, Bulgaria

Outline

- Formulation of importance sampling problem.
- Background material on currently suggested adaptive importance sampling methods.
- The proposed methodology
- Numerical example taken from recent work by Asmussen, Blanchet, Juneja, and Rojas-Nandayapa — *Tail probabilities of sums of correlated lognormals*.
- Conclusions

Problem formulation

- The problem is to estimate high-dimensional integrals of the form

$$\ell = \int f(\mathbf{x}) H(\mathbf{x}) \, \mathrm{d}\mathbf{x} = \mathbb{E}_f H(\mathbf{X})$$

- The function $H : \mathbb{R}^d \to \mathbb{R}$ and $\mathbf{X}$ is a d-dimensional random variable with pdf f.
- For discrete counting or combinatorial problems we simply replace the integration by summation.
- How do we estimate such integrals via Monte Carlo?

Existing methods

- Importance sampling: Let g be an importance sampling density such that $g(\mathbf{x}) = 0 \;\Rightarrow\; H(\mathbf{x})\, f(\mathbf{x}) = 0$ for all $\mathbf{x}$.
- Generate $\mathbf{X}_1, \ldots, \mathbf{X}_m \overset{\text{iid}}{\sim} g$, then an unbiased estimator of ℓ is

$$\widehat{\ell} = \frac{1}{m} \sum_{k=1}^{m} Z_k \quad \text{with} \quad Z_k = H(\mathbf{X}_k)\, \frac{f(\mathbf{X}_k)}{g(\mathbf{X}_k)} \; .$$

- The minimum variance importance sampling density is

$$\pi(\mathbf{x}) = \frac{H(\mathbf{x})\, f(\mathbf{x})}{\ell} \; , \quad (H(\mathbf{x}) \geqslant 0) \; ,$$

which depends on ℓ and is therefore not useful.

Existing importance sampling ideas

- Existing methods for selecting the density g assume that g is part of a parametric family: $\{g(\cdot;\boldsymbol{\eta})\}$.
- The objective is to select the parameter $\boldsymbol{\eta}$ so that g is as "close" to the optimal density π as possible.
- The closeness between π and g is measured by the ϕ-divergence distance

$$\int \pi(\mathbf{x})\, \phi\left(\frac{g(\mathbf{x};\boldsymbol{\eta})}{\pi(\mathbf{x})}\right) \mathrm{d}\mathbf{x}\ , \tag{1}$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}$ is twice continuously differentiable, and $\phi(1) = 0,\ \phi''(x) > 0$, for all $x > 0$.

Variance Minimal method

Variance Minimal (VM) method:

- Method equivalent to minimizing the ϕ-divergence with $\phi(z) = 1/z$.
- The minimization of the resulting ϕ-divergence

$$\underset{\boldsymbol{\eta}}{\operatorname{argmin}} \int \frac{\pi^2(\mathbf{x})}{g(\mathbf{x};\boldsymbol{\eta})} \mathrm{d}\mathbf{x}$$

  is highly nonlinear ☹
- The ϕ-divergence has to be estimated — we have a noisy nonlinear optimization problem ☹

Cross Entropy method

Cross Entropy method:

- Method equivalent to minimizing the ϕ-divergence with $\phi(z) = -\ln(z)$.
- The minimization of the Kullback-Leibler distance

$$\underset{\boldsymbol{\eta}}{\operatorname{argmin}} - \int \pi(\mathbf{x}) \ln(g(\mathbf{x}; \boldsymbol{\eta})/\pi(\mathbf{x}))\mathrm{d}\mathbf{x}$$

  is similar to likelihood maximization and we thus frequently have analytical solutions to the optimization problem ☺
- The Kullback-Leibler distance still has to be estimated ☹

Shortcomings

Both Cross-Entropy and Variance Minimization methods suffer from the following ☹:

- Their performance is limited by how well a simple and rigid parametric density can approximate the optimal π. Ideally we would like a flexible non-parametric model for the importance sampling density, but this is often impossible.
- The VM method always requires non-linear non-convex optimization and the Cross Entropy method is as simple as likelihood maximization can be.

MCMC methods for estimation

The Bayesian community has alternatives to parametric importance sampling that use MCMC:

- Chib's method
- Bridge sampling
- Path sampling
- Equi-Energy sampling
- Gelfand-Dey Method

The most popular and efficient is Chib's method and its variants. However, even Chib's approach suffers from the following drawbacks.

MCMC problems

- The estimators are biased estimators, because the chain almost never starts in stationarity.
- Difficulty in computing empirical and asymptotic error estimates, because MCMC does not generate iid samples.
- Chib's estimator relies on the output of multiple different chains.

Is there a way to draw on the strength of both MCMC and importance sampling?

Markov chain importance sampling

Similar to the cross entropy and variance minimization methods, the MCIS method consists of two stages:

- **Markov Chain (MC)** stage, in which we construct an ergodic estimator $\widehat{\pi}$ of the minimum variance importance sampling density π.
- **Importance Sampling (IS)** stage, in which we use $\widehat{\pi}$ as an importance sampling density to estimate ℓ.

There are many different ways of constructing a model-free estimator of π — e.g., standard kernel density estimation (poor convergence). Here we only explore one way related to the Gibbs sampler.

MCIS with Gibbs

- Suppose that we have used an MCMC sampler to generate the population

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{approx}{\sim} \pi(\mathbf{x}), \quad \mathbf{X}_i = (X_{i,1}, \ldots, X_{i,d}) \; .$$

- We can construct the nonparametric estimator of π using the Gibbs transition kernel:

$$\widehat{\pi}(\mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \kappa(\mathbf{y} \,|\, \mathbf{X}_i)$$

$$\kappa(\mathbf{y} \,|\, \mathbf{X}) \overset{\text{def}}{=} \prod_{j=1}^{d} \pi(y_j \,|\, y_1, \ldots, y_{j-1}, X_{j+1}, \ldots, X_d) \; .$$

MCIS vs importance sampling

The MCIS estimator is

$$\widehat{\ell} = \frac{1}{m} \sum_{k=1}^{m} \frac{|H(\mathbf{Y}_k)|\, f(\mathbf{Y}_k)}{\widehat{\pi}(\mathbf{Y}_k)} \,,$$

where $\mathbf{Y}_1, \ldots, \mathbf{Y}_m \overset{iid}{\sim} \widehat{\pi}$. Advantages and disadvantages of the MCIS approach:

- $\mathbb{P}(\lim_{n\to\infty} \widehat{\pi}(\mathbf{x}) = \pi(\mathbf{x})) = 1$ for all $\mathbf{x}$.
- No need to solve a ϕ-divergence optimization problem
- If n is large, then evaluation of $\widehat{\pi}$ may be costly.
- MCIS estimator, unlike Chib's, is unbiased and iid sampling allows for standard estimation error bands.

Sums of correlated lognormals

Consider the estimation of the rare-event probability

$$\ell = \mathbb{P}(\mathrm{e}^{X_1} + \cdots + \mathrm{e}^{X_d} \geqslant \gamma) = \int f(\mathbf{x})\, \mathrm{I}\{S(\mathbf{x}) \geqslant \gamma\}\, \mathrm{d}\mathbf{x}\ ,$$

where:

- $\mathbf{X} = (X_1, \ldots, X_d)$ and $\mathbf{X} \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$.
- $S(\mathbf{x}) = \mathrm{e}^{x_1} + \cdots + \mathrm{e}^{x_d}$.
- f is the density of $\mathsf{N}(\boldsymbol{\mu}, \Sigma)$ with associated precision matrix $\Lambda = \Sigma^{-1}$.
- We use the notation $\Lambda = (\Lambda_{i,j})$ and $\Sigma = (\Sigma_{i,j})$.

Conditional densities needed for $\widehat{\pi}$

We need the conditional densities of the optimal importance sampling pdf $\pi(\mathbf{y}) = f(\mathbf{y})\mathrm{I}\{S(\mathbf{x}) \geqslant \gamma\}/\ell$:

$$\pi(y_i \,|\, \mathbf{y}_{-i}) \propto \begin{cases} f(y_i \,|\, \mathbf{y}_{-i}) & \text{if } \sum_{j\neq i} \mathsf{e}^{y_j} \geqslant \gamma \\ f(y_i \,|\, \mathbf{y}_{-i})\, \mathrm{I}\Big\{ y_i \geqslant \ln(\gamma - \sum_{j\neq i} \mathsf{e}^{y_j}) \Big\} & \text{if } \sum_{j\neq i} \mathsf{e}^{y_j} < \gamma \end{cases} ,$$

where by standard properties of the multivariate normal density $f(y_i \,|\, \mathbf{y}_{-i})$ is normal density with mean

$$\mu_i + \Lambda_{i,i}^{-1} \sum_{j\neq i} \Lambda_{i,j}(\mu_j - y_j) \quad \text{and variance} \quad \Lambda_{i,i}^{-1} \ .$$

Numerical setup

- Compare with importance sampling vanishing relative error estimator (ISVE) and the cross entropy vanishing relative error estimator (CEVE) of *Asmussen, Blanchet, Juneja, Rojas-Nandayapa*, Efficient simulation of tail probabilities of sums of correlated lognormals, Ann. Oper. Res. (2009)
- Both of these estimators decompose the probability $\ell(\gamma)$ into two parts:

$$\ell(\gamma) = \mathbb{P}(\max_i X_i \geqslant \gamma) + \mathbb{P}(S(\mathbf{X}) \geqslant \gamma,\ \max_i X_i \leqslant \gamma).$$

- The first (dominant) term and the second (residual) term are estimated by two different importance sampling estimators that ensure the strong efficiency of the sum.

Numerical setup I

- The first term is asymptotically dominant in the sense that

$$\lim_{\gamma\to\infty} \frac{\mathbb{P}(\max_i X_i \geqslant \gamma)}{\mathbb{P}(S(\mathbf{X}) \geqslant \gamma,\ \max_i X_i \leqslant \gamma)} = 0\ .$$

- We compute ℓ for various values of the common correlation coefficient

$$\varrho = \frac{\Sigma_{i,j}}{\sqrt{\Sigma_{i,i}\Sigma_{j,j}}}.$$

- $d = 10,\ \mu_i = i - 10,\ \sigma_i^2 = i\ (i = 1, \ldots, d),\ \gamma = 5 \times 10^5$.
- We used the MCIS estimator with $m = 5 \times 10^5$ and a Markov chain sample $n = 80$ obtained using splitting.

Numerical results I

ϱ	MCIS est. $\widehat{\ell}$	relative error %		
		MCIS	CEVE	ISVE
0	1.795092×10^{-5}	0.0092	0.0063	0.0069
0.4	1.8077×10^{-5}	0.093	0.17	0.23
0.7	1.9014×10^{-5}	0.04	0.65	2.85
0.9	2.0735×10^{-5}	0.068	0.63	2.80
0.93	2.0997×10^{-5}	0.17	3.5	4.59
0.95	2.1412×10^{-5}	0.11	6	8.09
0.99	2.1882×10^{-5}	0.29	15	4.35

Numerical setup II

- Both the ISVE and CEVE estimators are strongly efficient, so we expect that these estimators will eventually outperform the MCIS estimator.
- This intuition is confirmed in the next table describing 14 cases with increasing values of the threshold γ.
- We use the same algorithmic and problem parameters, except that $\varrho = 0.9$ in all cases and the threshold parameter depends on the case number c according to the formula:

$$\gamma = 5 \times 10^{c+3}, \quad c = 1, \ldots, 14 \; .$$

Numerical results II

		relative error %		efficiency	
case $c =$	ISVE estimate	MCIS	ISVE	MCIS	ISVE
1	3.9865×10^{-4}	0.049	1.6	25	1500
2	2.0802×10^{-5}	0.067	4.3	75	10000
3	6.4385×10^{-7}	0.077	5.5	64	22000
4	1.2039×10^{-8}	0.041	3.0	21	5000
5	$1.3468e \times 10^{-10}$	0.034	6.1	12	20000
6	8.9791×10^{-13}	0.036	7.6	17	30000
7	3.5899×10^{-15}	0.043	0.014	27	0.11
8	8.5302×10^{-18}	0.079	0.029	81	0.50
9	1.2082×10^{-20}	0.025	0.024	7	0.32
10	1.0148×10^{-23}	0.042	0.00064	28	0.00021
11	5.0428×10^{-27}	0.042	0.00030	27	4.6×10^{-5}
12	1.4898×10^{-30}	0.024	0.00014	7	1.0×10^{-5}
13	2.5961×10^{-34}	0.023	3.87×10^{-15}	8.2	7.7×10^{-27}
14	2.6754×10^{-38}	0.012	4.22×10^{-13}	2.6	9.1×10^{-23}

L^2 properties

Assume that:

- $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{\text{iid}}{\sim} \kappa_{t-1}(\mathbf{x} \,|\, \boldsymbol{\theta})$, where κ_{t-1} is the $(t-1)$ step transition density.
- κ is the transition density of systematic Gibbs sampler.
- $\iint \kappa^2(\mathbf{y} \,|\, \mathbf{x}) \frac{\pi(\mathbf{x})}{\pi(\mathbf{y})} \mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} < \infty$. Warning: this can be difficult to verify ☹
- All states of the chain can be accessed using a single step of the chain (irreducability condition).

Then we have the following result for $\widehat{\ell} = \frac{f(\mathbf{Y})H(\mathbf{Y})}{\widehat{\pi}(\mathbf{Y})}$.

L^2 theoretical properties

- We have the following bound on the Neymann χ^2 distance:

$$\mathbb{E}\left[\frac{(\widehat{\ell}-\ell)^2}{\widehat{\ell}\,\ell}\right] = \underbrace{\left(-1+\int \frac{\kappa_t^2(\mathbf{y}\,|\,\boldsymbol{\theta})}{\pi(\mathbf{y})}\mathrm{d}\mathbf{y}\right)}_{\chi^2 \text{ component}} + \underbrace{\frac{1}{n}\int \frac{\mathbb{E}\kappa^2(\mathbf{y}\,|\,\mathbf{X})-\kappa_t^2(\mathbf{y}\,|\,\boldsymbol{\theta})}{\pi(\mathbf{y})}\mathrm{d}\mathbf{y}}_{\text{variance component}}$$

$$\leqslant V(\boldsymbol{\theta})\mathrm{e}^{-rt}+\mathcal{O}(1/n)\ ,$$

where V is a positive Liapunov function and $r>0$ is a constant (the geometric rate of convergence of the Gibbs sampler).

- Note that the above is NOT the relative error, because the denominator has $\widehat{\ell}$, instead of ℓ.

Conclusions

- We have presented a new method for integration that combines MCMC with importance sampling.
- We argue that the method is preferable to methods based purely on MCMC or importance sampling.
- The method dispenses with the traditional parametric modeling used in the design of importance sampling densities ☺
- Unlike MCMC based methods, the proposed method provides unbiased estimators and estimation of relative error and confidence bands is straightforward ☺

Some interesting links

- Can we use large deviations theory to sample *exactly* from the minimum variance density, instead of using MCMC?
- Idea related to empirical likelihood and Rao-Blackwellization in classical statistics. Essentially we construct an estimator based on empirical likelihood arguments.

thank you for your attention