Rare-event probability estimation and convex programming

Z. I. Botev

Problem formulation

- We consider the problem of estimating rare-event probabilities of the form

$$\ell = \mathbb{P}(S(\mathbf{X}) > \gamma), \quad \mathbf{X} = (X_1, \ldots, X_d),$$

  where $S : \mathbb{R}^d \to \mathbb{R}$ and $X_1, \ldots, X_d$ are random variables with joint density $f(\mathbf{x})$. Almost any high-dimensional integration problem can be cast into this framework: option pricing, insurance value at risk, ABC, Ising model in physics, queuing models, counting problems in operations research and CS.
- Dichotomy: Frequently, one can easily simulate the rare event, but it is not clear how one can use the simulated data to estimate ℓ.

Bird's eye view of existing methods

- importance sampling: state-dependent and state-independent; no need to simulate the rare-event under the original probability law in order to estimate efficiently the probability of its occurrence.
- adaptive importance sampling: cross entropy method
- conditioning
- splitting: splitting method typically yields exact or approximate realizations of the rare-event

With the exception of splitting, all treat the problem of simulating the rare-event and estimating the rare-event probability as essentially separate problems.

Standard importance sampling

- We propose to estimate the quantity ℓ using maximum likelihood methods, where the data upon which the likelihood function depends is generated from computer simulation of the rare-event under the original probability law.
- To introduce the idea it is convenient to think of ℓ as a normalization constant ℓ_s of the conditional density

$$f_s(\mathbf{x}) = \frac{f(\mathbf{x})\mathbb{I}\{S(\mathbf{x}) > \gamma\}}{\ell_s} = \frac{w_s(\mathbf{x})}{\ell_s} .$$

The typical importance sampling scheme

- Let $f_1(\mathbf{x}) = w_1(\mathbf{x})/\ell_1$ be another density whose normalizing constant ℓ_1 is known and $\{\mathbf{x} : f_1(\mathbf{x}) > 0\} \supseteq \{\mathbf{x} : f_s(\mathbf{x}) > 0\}$.
- Then, the natural estimator is

$$\ell_s^* = \frac{1}{n} \sum_{j=1}^{n} \frac{f(\mathbf{X}_j)\mathbb{I}\{S(\mathbf{X}_j) > \gamma\}}{f_1(\mathbf{X}_j)}, \quad \mathbf{X}_1, \ldots, \mathbf{X}_n \overset{\text{iid}}{\sim} f_1 \; .$$

- For good performance we need the tails of f_1 to be at least as heavy as the tails of f_s.
- Frequently, the choice of f_1 is dictated by asymptotic analysis of $\ell(\gamma)$ as $\gamma \uparrow \infty$.

The typical importance sampling scheme

- It is well known that the zero-variance importance sampling density for estimating ℓ_s is the conditional pdf f_s. So we may try to use f_s itself as an importance sampling density.
- We consider the estimator

$$\widehat{\ell}_s = \frac{1}{n}\sum_{j=1}^{n} \frac{w_s(\mathbf{X}_j)}{\lambda_1 f_1(\mathbf{X}_j) + \lambda_s \underbrace{w_s(\mathbf{X}_j)/\widehat{\ell}_s}_{\approx f_s}}, \qquad \mathbf{X}_1,\ldots,\mathbf{X}_n \stackrel{\text{iid}}{\sim} \bar{f},$$

where the normalizing constant ℓ_s on the right is replaced with $\widehat{\ell}_s$, giving rise to a nonlinear equation for $\widehat{\ell}_s$.

Comparison of estimators

- Compare now the traditional importance sampling estimator with the proposed solve-the-equation-type.
- For the solve-the-equation-type the support and tail restrictions on f_1 are no longer necessary for good performance.
- The only requirement is that $\{\mathbf{x} : f_1(\mathbf{x}) \times f_s(\mathbf{x}) > 0\} \neq \emptyset$, that is, the supports of f_1 and f_s overlap.
- This is an example of using computer simulated data of the rare-event to obtain an M-type estimator of ℓ. Next, we try to generalize.

- Suppose we are given the sequence of densities

$$f_t(\mathbf{x}) = \frac{w_t(\mathbf{x})}{\ell_t} = \frac{f(\mathbf{x})H_t(\mathbf{x})}{\ell_t}, \quad t = 1, \ldots, s \,,$$

  where f is a known density, $\{H_t\}$ are known functions, and ℓ_t are probabilities acting as normalizing constants to $\{w_t\}$.
- We are interested in estimating $\ell_{k'} = \mathbb{E}_f H_{k'}(\mathbf{X})$ for some k'.
- We assume that for at least one f_t, say f_1, the corresponding normalizing constant ℓ_1 is known, and, without loss of generality, equal to unity.
- We call f_1 a reference density.

Connectivity

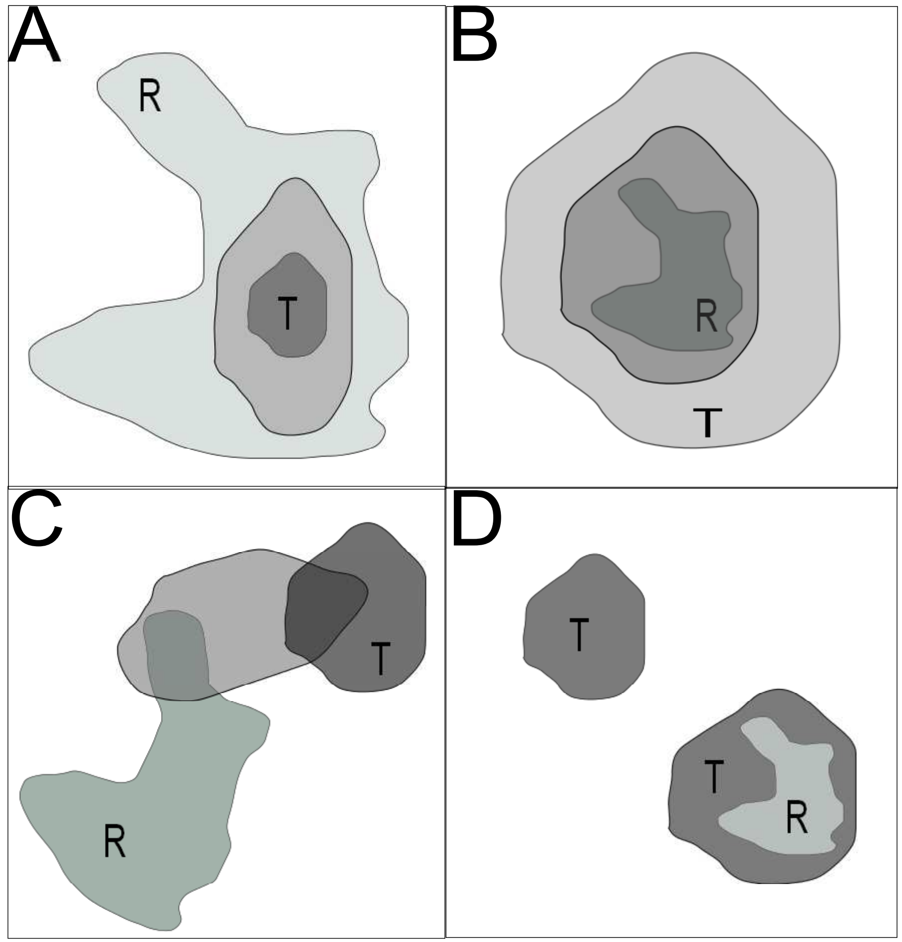
- To proceed, suppose we are given a graph with s nodes and an edge between nodes i and j if and only if

$$\mathbb{E}_f \mathbb{I}\{H_i(\mathbf{X}) > 0\} \times \mathbb{I}\{H_j(\mathbf{X}) > 0\} > 0. \quad (1)$$

- We assume that there exists a path between any two nodes.
- We call the condition (**??**) on the supports of $\{f_t\}$ *Vardi's connectivity condition*.
- Assume we have the iid sample

$$\mathbf{X}_{t,1}, \ldots, \mathbf{X}_{t,n_t} \overset{\text{iid}}{\sim} f_t(\mathbf{x}), \quad t = 1, \ldots, s \; .$$

Connectivity



Mixture model

- Conceptually, same as sampling $n = n_1 + \cdots + n_s$ variables with stratification from mixture

$$\bar{f}(\mathbf{x}) = \frac{1}{n}\sum_{t=1}^{s} n_t f_t(\mathbf{x}) = \sum_{t=1}^{s} \lambda_t f_t(\mathbf{x}), \quad \lambda_t \stackrel{\text{def}}{=} n_t/n \ .$$

- Let the pooled sample be denoted via

$$\mathbf{X}_1, \ldots, \mathbf{X}_n,$$

where the first n_1 samples are outcomes from f_1, the next n_2 are samples from f_2, and so on.

- Define the vector of parameters

$$\mathbf{z} = (z_1, \ldots, z_n) = (-\ln(1/\lambda_1), -\ln(\ell_2/\lambda_2), \ldots, -\ln(\ell_s/\lambda_s)).$$

Empirical Likelihood Estimator

Now consider the likelihood of the observed data $\mathbf{X}_{t,1}, \ldots, \mathbf{X}_{t,n_t},\ t = 1, \ldots, s$ as a function of $\mathbf{z}$:

$$\prod_{k=1}^{s} \prod_{j=1}^{n_k} f_k(\mathbf{X}_{k,j}) = \prod_{k=1}^{s} \prod_{j=1}^{n_k} \frac{w_k(\mathbf{X}_{k,j})}{\lambda_k \mathrm{e}^{-z_k}}$$

$$= \prod_{k=1}^{s} \prod_{j=1}^{n_k} \frac{w_k(\mathbf{X}_{k,j})}{\lambda_k \mathrm{e}^{-z_k} \bar{f}(\mathbf{X}_{k,j})} \times \prod_{k=1}^{s} \prod_{j=1}^{n_k} \bar{f}(\mathbf{X}_{k,j}) \ .$$

Empirical Likelihood Estimator

The last yields the partial log-likelihood as a function of $\mathbf{z}$:

$$\ln \prod_{k=1}^{s} \prod_{j=1}^{n_k} \frac{w_k(\mathbf{X}_{k,j})}{\lambda_k \mathsf{e}^{-z_k} \bar{f}(\mathbf{X}_{k,j})} =$$

$$= \sum_{k=1}^{s} \sum_{j=1}^{n_k} \ln(w_k(\mathbf{X}_{k,j})/\lambda_k) + z_k - \ln(\bar{f}(\mathbf{X}_{k,j}))$$

$$= \text{const.} - \sum_{j=1}^{n} \ln(\bar{f}(\mathbf{X}_j)) + \sum_{k=1}^{s} n_k z_k$$

$$= \text{const.} - \sum_{j=1}^{n} \ln\left(\sum_{k=1}^{s} w_k(\mathbf{X}_j)\, \mathsf{e}^{z_k}\right) + n \sum_{k=1}^{s} \lambda_k z_k \ .$$

Empirical Likelihood Estimator

- Under the connectivity condition of Vardi and iid assumption on the sample, it can be shown that the maximum of the partial log-likelihood is the same as the maximum of the complete likelihood.
- In other words, the unique nonparametric maximum likelihood estimate of $\mathbf{z}$ (and hence of ℓ) solves the almost surely convex optimization program (with $z_1 = \ln \lambda_1$ fixed)

$$\widehat{\mathbf{z}} = \operatorname*{argmin}_{\mathbf{z}} \widehat{D}(\mathbf{z})$$

$$\widehat{D}(\mathbf{z}) \stackrel{\text{def}}{=} \sum_{j=1}^{n} \ln \left(\sum_{k=1}^{s} w_k(\mathbf{X}_j)\, \mathrm{e}^{z_k} \right) - \sum_{k=1}^{s} n_k z_k \;, \tag{2}$$

Almost surely??

- Define the matrix

$$G_{i,j} = \mathbb{E}_f \mathbb{I}\{H_i(\mathbf{X}) > 0\} \times \mathbb{I}\{H_j(\mathbf{X}) > 0\}$$

and its empirical counterpart

$$\widehat{G}_{i,j} = \frac{1}{n} \sum_k \mathbb{I}\{H_i(\mathbf{X}_k) > 0\} \times \mathbb{I}\{H_j(\mathbf{X}_k) > 0\}$$

- Under the connectivity condition, matrix G is irreducible, that is if for any pair (i, j), we have $\widehat{G}^k_{i,j} > 0$ for some $k = 1, 2, \ldots,$, and since $\widehat{G} \to G$ almost surely, with probability one the function $\widehat{D}(\mathbf{z})$ is a strictly convex and the maximum likelihood program has a unique solution.

M-estimator & Method of Moments

- To execute the optimization, we compute the $(s-1)\times 1$ gradient $\nabla\widehat{D}(\mathbf{z})$, which is a vector with components:

$$[\nabla\widehat{D}]_t(\mathbf{z}) = \sum_{j=1}^{n} \frac{w_t(\mathbf{X}_j)\,\mathrm{e}^{z_t}}{\sum_{k=1}^{s} w_k(\mathbf{X}_j)\,\mathrm{e}^{z_k}} - n_t, \qquad t = 2,\ldots,s\ .$$

- Using the prior information that $\ell_1 = 1$ or $z_1 = \ln(\lambda_1)$, these estimating equations are equivalent to solving the $s-1$ dimensional system for the unknown $z_2,\ldots,z_s$:

$$\frac{1}{n}\sum_{j=1}^{n} \frac{w_t(\mathbf{X}_j)\,\mathrm{e}^{\widehat{z}_t}}{\sum_{k=1}^{s} w_k(\mathbf{X}_j)\,\mathrm{e}^{\widehat{z}_k}} = \lambda_t, \qquad t = 2,\ldots,s\ .$$

Moment-matching master equation

- The last bit is equivalent to:

$$\widehat{\ell}_t = \sum_{j=1}^{n} \frac{A_{t,j}}{\sum_{k=1}^{s} A_{k,j}\, n_k / \widehat{\ell}_k}, \qquad A_{t,j} = H_t(\mathbf{X}_j), \qquad t = 2, \ldots, s\ . \tag{3}$$

- It is sometimes easier to directly minimize $\widehat{D}$, instead of solving the nonlinear system.
- The system can be solved using Jacobi/Gauss-Seidel type iteration.
- The process is similar to finding eigenvalues via power iteration.

Jacobi/Gauss-Seidel iteration algorithm

Require: Matrix A and initial starting point

$\boldsymbol{\ell} = (\ell_1, \ldots, \ell_s) = (1, \ldots, 1)$

Set $\varepsilon = \infty$ and $\boldsymbol{\ell}^* \leftarrow \boldsymbol{\ell}$

while $\varepsilon > 10^{-10}$ **do**

for $i = 2, \ldots, s$ **do**

$$\ell_i \leftarrow \sum_{j=1}^{n} \frac{A_{i,j}}{\sum_{k=1}^{s} A_{k,j}\, n_k / \ell_k^*}$$

$\varepsilon \leftarrow \max_i \frac{|\ell_i - \ell_i^*|}{\ell_i}$

Set $\boldsymbol{\ell}^* \leftarrow \boldsymbol{\ell}$

return The vector of estimated probabilities $\widehat{\boldsymbol{\ell}} \leftarrow \boldsymbol{\ell}$.

Error estimates

- Using the properties of maximum likelihood estimators, one can derive the asymptotic covariance matrix of $(\widehat{\ell}_2, \ldots, \widehat{\ell}_s)$.
- First, define the $s \times s$ matrix $O^{\#}$ with entries

$$O^{\#}_{i,j} = \int \frac{f_i(\mathbf{x}) f_j(\mathbf{x})}{\sum_{k=1}^{s} \lambda_k f_k(\mathbf{x})} \mathrm{d}\mathbf{x} = \mathbb{E}_{\bar{f}} \left[\frac{w_i(\mathbf{X})/\ell_i \; w_j(\mathbf{X})/\ell_j}{\left(\sum_{k=1}^{s} w_k(\mathbf{X}) \lambda_k / \ell_k\right)^2} \right]$$

$$= \mathbb{E}_{\bar{f}} \left[\frac{\mathbb{I}\{S_i(\mathbf{X}) > \gamma\}/\ell_i \; \mathbb{I}\{S_j(\mathbf{X}) > \gamma\}/\ell_j}{\left(\sum_{k=1}^{s} \mathbb{I}\{S_k(\mathbf{X}) > \gamma\} \, \lambda_k/\ell_k\right)^2} \right],$$

$$i, j = 1, \ldots, s \; .$$

This matrix has an obvious plug-in estimator.

Conclusions and Future work

Next, define the matrices (with the specified dimension)

$$J = \begin{pmatrix} \mathbf{0} \\ I_{(s-1)\times(s-1)} \end{pmatrix}, \qquad s \times (s-1)$$

$$O = J^\top O^{\#} J, \qquad (s-1) \times (s-1) \quad \text{(lower right submatrix of } O^{\#}\text{)}$$

$$\Lambda = \mathrm{diag}(\lambda_2, \ldots, \lambda_s), \qquad (s-1) \times (s-1)$$

$$L = \mathrm{diag}(\ell_2, \ldots, \ell_s), \qquad (s-1) \times (s-1)\ .$$

Asymptotic distribution

Then, we have for large $n \uparrow \infty$ and fixed $\lambda_i = \frac{n_i}{n_1 + \cdots + n_s} > 0$ (and assuming $\ell_1 = 1$)

$$\sqrt{n} J^\top (\widehat{\boldsymbol{\ell}} - \boldsymbol{\ell}) \stackrel{d}{\to} \mathsf{N}\left(\mathbf{0}, L(O^{-1} - \Lambda)^{-1} L\right) .$$

Example

Consider the estimation of the rare-event probability

$$\ell_2 = \mathbb{P}(\mathrm{e}^{X_1} + \cdots + \mathrm{e}^{X_d} \geqslant \gamma),$$

which is the normalizing constant of the density

$$f_2(\mathbf{x}) = \frac{f(\mathbf{x})\,\mathbb{I}\{S_2(\mathbf{x}) \geqslant \gamma\}}{\ell_2}\ , \quad \mathbf{x} = (x_1, \ldots, x_d),$$

where $S_2(\mathbf{x}) = \mathrm{e}^{x_1} + \cdots + \mathrm{e}^{x_d}$, and f is the density of the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma = (\Sigma_{i,j})$ with

$$\frac{\Sigma_{i,j}}{\sqrt{\Sigma_{i,i}\Sigma_{j,j}}} = \varrho, \quad \text{for all } i \neq j\ ,$$

Reference density

- Estimate ℓ_2 via the M-estimator with $s = 2$ and reference density ($(X_1, \ldots, X_d) \sim \mathsf{N}(\boldsymbol{\mu}, \Sigma)$)

$$f_1(\mathbf{x}) = \frac{f(\mathbf{x}) \sum_{j=1}^{d} \mathbb{I}\{\mathsf{e}^{x_i} > \gamma\}}{\ell_1}, \qquad \ell_1 \stackrel{\text{def}}{=} \sum_{j=1}^{d} \mathbb{P}\left(\mathsf{e}^{X_i} > \gamma\right) .$$

- Sampling iid copies from the reference density using the mixture representation

$$f_1(\mathbf{x}) = \sum_{j=1}^{d} \frac{\mathbb{P}(X_j > \ln \gamma)}{\ell_1} \frac{f(\mathbf{x}) \mathbb{I}\{x_j > \ln \gamma\}}{\mathbb{P}(X_j > \ln \gamma)}.$$

Reference density

- The reference density has the property that $\ell_2 \downarrow \ell_1$ as $\gamma \uparrow \infty$, which makes is possible to show that the M-estimator has vanishing relative error properties.
- Note, however, that the reference density cannot be used as an importance sampling pdf in a traditional scheme due to its support.
- Moreover, the behavior of the reference density does not capture the effect of the correlation coefficient ϱ. The parameter ϱ does not appear anywhere.
- This illustrates the advantages of the proposed method.

Moment-matching master equation

- $$\ell_2 = \sum_{j=1}^{n} \frac{w_2(\mathbf{X}_j)}{n_1\, w_1(\mathbf{X}_j)/\ell_1 + n_2\, w_2(\mathbf{X}_j)/\ell_2}$$
$$= \frac{p_0}{\frac{0\,n_1}{\ell_1} + \frac{n_2}{\ell_2}} + \frac{p_1}{\frac{n_1}{\ell_1} + \frac{n_2}{\ell_2}} + \frac{p_2}{\frac{2\,n_1}{\ell_1} + \frac{n_2}{\ell_2}} + \cdots + \frac{p_d}{\frac{d\,n_1}{\ell_1} + \frac{n_2}{\ell_2}},$$

  where p_k is the number of $\mathbf{X}_j$'s, which yield $\sum_{i=1}^{d} \mathbb{I}\{x_i > \ln\gamma\} = k$.

- Hence, our estimator $\widehat{\ell_2}$ solves the equation:

$$p_0 + \frac{p_1}{\ell_2 \frac{n_1}{n_2\,\ell_1} + 1} + \frac{p_2}{\ell_2 \frac{2\,n_1}{n_2\,\ell_1} + 1} + \cdots + \frac{p_d}{\ell_2 \frac{d\,n_1}{n_2\,\ell_1} + 1} = n_2\ .$$

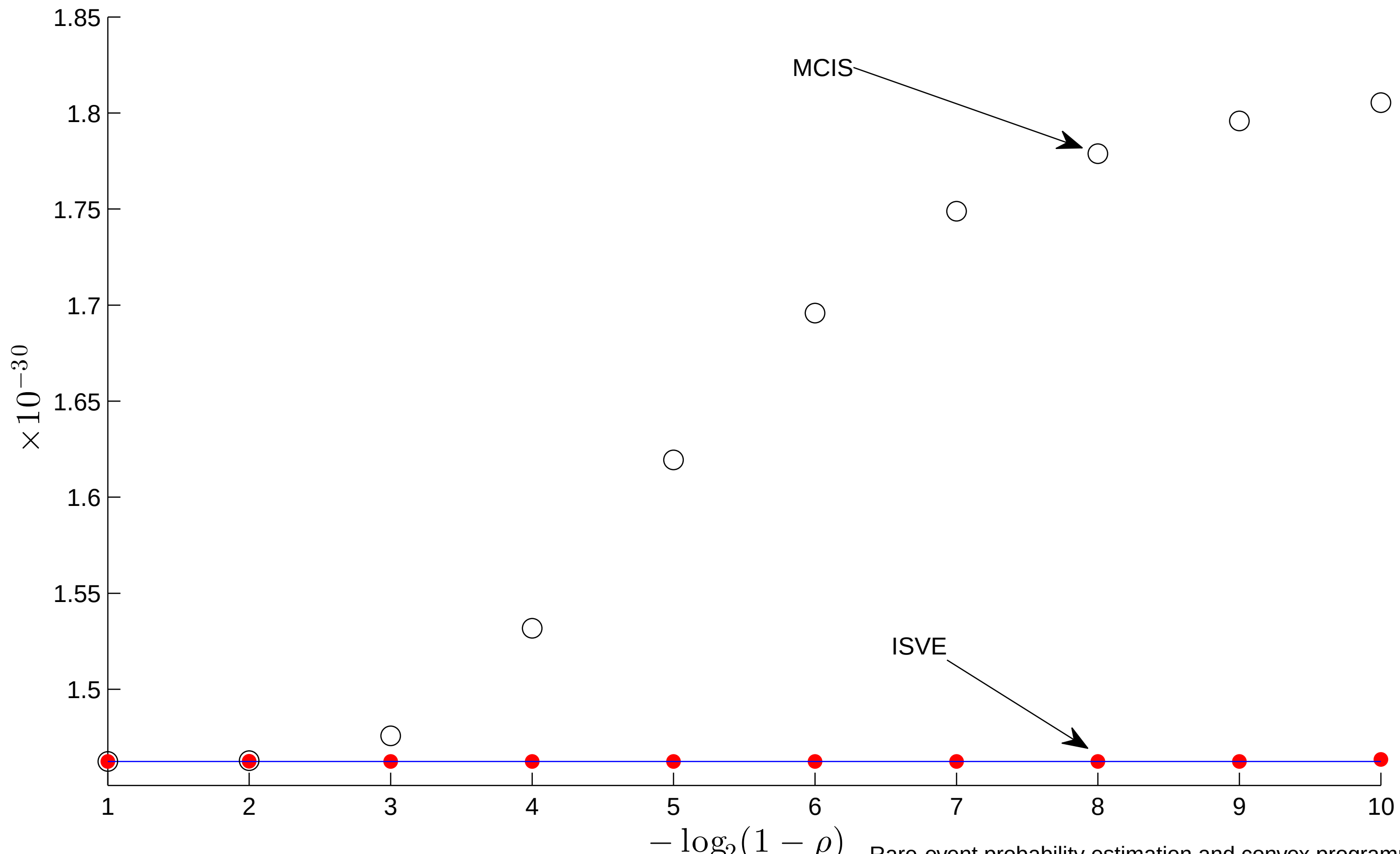
Numerical results

Empirical performance of M-estimator and ISVE algorithms for various values of the threshold parameter $\gamma = 5 \times 10^{c+3}$, $c = 1, \ldots, 14$ with $\varrho = 0.999$. Both algorithms use a sample size of $n = n_1 + n_2 = 5 \times 10^5$.

				relative error %		WNRV	
γ	asym. approx.	M-estim.	ISVE estim.	M-estim.	ISVE	M-estim.	ISVE
5×10^4	0.000355	0.000409	0.000406	0.23	1.71	0.00044	15248
5×10^5	1.794×10^{-5}	2.212×10^{-5}	2.177×10^{-5}	0.23	3.09	0.00043	50267
5×10^6	5.586×10^{-7}	7.156×10^{-7}	6.807×10^{-7}	0.23	5.32	0.00042	1.4×10^5
5×10^7	1.057×10^{-8}	1.384×10^{-8}	1.444×10^{-8}	0.23	11.74	0.00042	7.2×10^5
5×10^8	1.205×10^{-10}	1.590×10^{-10}	1.254×10^{-10}	0.23	2.35	0.00042	29064
5×10^9	8.230×10^{-13}	1.086×10^{-12}	3.781×10^{-12}	0.23	76.90	0.00040	3.13×10^7
5×10^{10}	3.347×10^{-15}	4.372×10^{-15}	3.346×10^{-15}	0.22	0.10	0.00040	56.12
5×10^{11}	8.087×10^{-18}	1.046×10^{-17}	8.083×10^{-18}	0.22	0.024	0.00039	2.99
5×10^{12}	1.158×10^{-20}	1.483×10^{-20}	1.158×10^{-20}	0.22	0.0018	0.00039	0.016
5×10^{13}	9.827×10^{-24}	1.245×10^{-23}	1.641×10^{-23}	0.22	40.12	0.00039	8.38×10^6
5×10^{14}	4.930×10^{-27}	6.170×10^{-27}	5.028×10^{-27}	0.22	1.94	0.00039	19790
5×10^{15}	1.462×10^{-30}	1.804×10^{-30}	1.462×10^{-30}	0.22	0.00037	0.00038	0.00073
5×10^{16}	2.562×10^{-34}	3.123×10^{-34}	2.563×10^{-34}	0.22	0.00020	0.00038	0.00020
5×10^{17}	2.651×10^{-38}	3.198×10^{-38}	2.652×10^{-38}	0.22	0.00010	0.00037	5.21×10^{-5}

Effect of correlation on estimate

Effect of the correlation parameter $\varrho = 1 - 0.5^c,\ c = 1, \ldots, 10$ on the rare-event probability. The circles represent the MCIS estimates and the dots lying on the line represent the ISVE estimates. The line itself is the asymptotic approximation of the rare-event probability. Both M-estim. and ISVE use a sample size of $n = n_1 + n_2 = 5 \times 10^6$.



Effect of correlation on probability

Empirical performance of M-estim. and ISVE algorithms for various values of the threshold parameter $\gamma = 5 \times 10^{15}$ with $\varrho = 1 - 0.5^c \;\; c = 1, \ldots, 10$. The asymptotic approximation here is $\approx 1.462 \times 10^{-30}$. Both M-estim. and ISVE use a sample size of $n = n_1 + n_2 = 5 \times 10^6$.

			relative error %		WNRV	
ϱ	M-estim.	ISVE estim.	M-estim.	ISVE	M-estim.	ISVE
$1 - 0.5^1$	1.4624×10^{-30}	1.4624×10^{-30}	0.063	3.58×10^{-14}	0.00027	5.89×10^{-22}
$1 - 0.5^2$	1.4629×10^{-30}	1.4624×10^{-30}	0.063	1.00×10^{-5}	0.00027	4.64×10^{-5}
$1 - 0.5^3$	1.4758×10^{-30}	1.4624×10^{-30}	0.063	0.00018	0.00028	0.015
$1 - 0.5^4$	1.5318×10^{-30}	1.4624×10^{-30}	0.064	9.54×10^{-5}	0.00029	0.0042
$1 - 0.5^5$	1.6194×10^{-30}	1.4624×10^{-30}	0.066	0.00011	0.00031	0.0053
$1 - 0.5^6$	1.6958×10^{-30}	1.4624×10^{-30}	0.068	0.00021	0.00032	0.019
$1 - 0.5^7$	1.7489×10^{-30}	1.4743×10^{-30}	0.069	0.78	0.00033	2.8×10^5
$1 - 0.5^8$	1.7788×10^{-30}	1.4624×10^{-30}	0.069	0.00010	0.00033	0.0050
$1 - 0.5^9$	1.7959×10^{-30}	1.4624×10^{-30}	0.070	0.00011	0.00034	0.0054
$1 - 0.5^{10}$	1.8054×10^{-30}	1.4624×10^{-30}	0.070	0.00010	0.00034	0.0048

Conclusions and Future work

- Apply to counting problems and Bayesian model choice problems.
- Finding λ using optimal design of experiments theory?!
- Splitting is an example where the moment-matching system of equations is exactly solvable.

$$Thank\ \ you!$$