Adaptive Importance Sampling to Estimate Power Grid Reliability

Vasilii Novitskii
Moscow Institute of Physics and Technology
9 Institutskiy per., Dolgoprudny, Russia 141701
vasiliy.novitskiy@phystech.edu +7 985 210 50 42

Samal Kubentayeva
Moscow Institute of Physics and Technology
9 Institutskiy per., Dolgoprudny, Russia 141701
samal.kubentaeva@gmail.com +7 925 373 20 34

Alena Shilova
Skolkovo Institute of Science and Technology
Nobelya Ulitsa 3, Moscow, Russia 121205
alena.shilova@skolkovotech.ru +7 977 356 00 69

Yuri Maximov
Los Alamos National Laboratory
T-4 CNLS, Los Alamos, NM 87544, USA
yury@lanl.gov +7 917 581 64 45

Abstract

We consider importance sampling to estimate the probability of a union of J rare events. There are existing ways of addressing the problem like ALOE. However, it ignores the geometrical structure of the problem. That is why we propose adaptive importance sampling. We show how our approach helps to improve the performance on both synthetic and real data.

Introduction

Now renewable energy sources become and more popular. The most well-spread are solar panels and wind farms. They are economically beneficial and environmentally friendly. However, they bring fluctuations to power grid and cause failures. The main focus is to estimate the probability of power system failure in the most efficient and accurate way.

Model

The most common way to describe power system work is to use alternating current (AC) power flow model. In this work we use linearized direct current (DC) model as a simplification of AC model. It is good in describing large power grids and is quite accurate for high voltage regimes.

Our aim is to estimate the probability of failure

$$\mu = \mathsf{P}\left(x \in \bigcup_{j=1}^{J} H_j\right). \tag{1}$$

We use Gaussian distribution for modeling of random busses such as variable demand by users and variable production (for example, wind farms), $x \sim \mathcal{N}(\eta, \Sigma)$. We model failure as union of J events H_j, $j \in \{1, \ldots, J\}$, each event means that one linear condition is violated.

Now we give an overview of how random vector x and events H_j come from DC model. Our power system is treated as a directed graph with N nodes (each node is one bus) and M edges (each edge is one power line). The power production at bus i is p_i, negative values indicate power consumption. We consider 3 types of busses: N_F fixed busses, N_R random busses and one slack bus S. The power at all busses can be represented by the vector $p = \left(p_F^\mathsf{T}, p_R^\mathsf{T}, p_S\right)^\mathsf{T}$. In (1) x is equal to $p_R \in \mathbb{R}^{N_R}$.

According to DC model we must satisfy conditions:

$$p_F = \eta_F \ (\eta_F \in \mathbb{R}^{N_F} \text{ is a constant}), \tag{2}$$

$$\sum_i p_i = 0 \text{ (balance equation)}, \tag{3}$$

$$\underline{p_i^R} \le p_i^R \le \overline{p_i^R} \ (\underline{p_i^R}, \overline{p_i^R} \in \mathbb{R}^{N_R} \text{ are constants}), \tag{4}$$

$$\underline{p_S} \le p_S \le \overline{p_S} \ (\underline{p_S}, \overline{p_S} \in \mathbb{R}^1 \text{ are constants}), \tag{5}$$

$$\theta = B^+ p \ (B \text{ is a given laplacian matrix}), \tag{6}$$

$$\left| \ \theta_i - \theta_j \ \right| \le \overline{\theta_{ij}} \ (i \ne j, \overline{\theta_{ij}} \text{ are constants}). \tag{7}$$

So p_S and θ are expressed through p_R from (3) and (6). From (4), (5) and (7) we have $J = 2N_R + 2 + 2M$ linear conditions on p_R which can be violated. Each event $H_j = \left\{x \mid w_j^\mathsf{T} x \ge \tau_j\right\}$ denotes that corresponding condition is violated.

Method

With the help of linear transformation $x \leftarrow \Sigma^{-1/2}(x - \eta)$ random vector x becomes standard multivariate Gaussian vector: $x \sim \mathcal{N}(0, I)$.

In the case of rare events importance sampling (IS) estimate of μ from (1) is better than Monte-Carlo estimate because IS estimate has lower variance. Power system failure is a rare event (typically $\mu < 10^{-3}$) so we use IS technique.

Denote $H = \cup_{j=1}^{J} H_j$, $p(x)$ is the PDF of $\mathcal{N}(0, I)$, $q(x)$ is the PDF of biased distribution q. Then IS estimate of μ is

$$\widehat{\mu}_q = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{I}(x_i \in H) p(x_i)}{q(x_i)}, \ x_i \sim q, \tag{8}$$

$$\text{Var}\left(\widehat{\mu}_q\right) = \frac{1}{n}\left(\int \frac{\mathbb{I}(x \in H) p^2(x)}{q(x)} dx - \mu^2\right). \tag{9}$$

It is important that

$$q(x) > 0 \text{ wherever } \mathbb{I}(x \in H) p(x) \ne 0. \tag{10}$$

Let us denote: $H_j(x) = \mathbb{I}(x \in H_j)$, $H(x) = \sum_{j=1}^{J} H_j(x)$, $P = P(H_j)$.

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The method ALOE was introduced and analyzed in (Owen, Maximov, and Chertkov 2017):

$$q^{ALOE}(x) \equiv q_\alpha(x) = \sum_{j=1}^{J} \alpha_j^{ALOE} q_j^{ALOE}(x), \quad (11)$$

$$q_j^{ALOE}(x) = \frac{H_j(x)p(x)}{P_j}, \quad (12)$$

$$\alpha_j^{ALOE} = P_j \left(\sum_{k=1}^{J} P_k \right)^{-1}. \quad (13)$$

It was proved in (Owen, Maximov, and Chertkov 2017) that

$$\text{Var}\left(\widehat{\mu}^{ALOE}\right) \leq \frac{\mu^2}{n} \frac{J + J^{-1} - 2}{4}.$$

The ALOE estimate is efficent only if $S(x) = \sum_{j=1}^{J} H_j(x)$ has a uniform distribution on $\{1, \ldots, J\}$.

Let $\mathcal{P}$ denote polytope: $\mathcal{P} = \cap_{j=1}^{J} \overline{H_j}$. Let polytope $\mathcal{P}_{\theta,J,\tau}$ be a regular pyramid in 3-dimention space where θ is an angle between base face and side face and J denotes the number of faces, which is equal to the number of linear conditions H_j. We set $P_j = e^{-\tau}$, so $H_j = \left\{x \mid w_j^{\mathsf{T}} x \geq \tau\right\}$, τ is the same for all j,

$$w_j = \begin{cases} (0, 0, -1), & \text{if } j = 1, \\ (\cos\alpha_j \sin\theta, \sin\alpha_j \sin\theta, \cos\theta), & \text{if } j \geq 2, \end{cases}$$

where $\alpha_j = \frac{2\pi(j-1)}{J-1}$. If $j = 1$ then we have the base face of the pyramid, if $j \geq 2$ then we have the side face.

In the case the angle θ is sufficiently small if a point x_i is sampled from q_j^{ALOE} where $j \geq 2$, then $S(x)$ is equal to $(J-1)$ with probability close to 1 (if $J = 1000, \theta = 0.01, \tau = 4$ then from 10000 samples $S(x) = J - 1$ in 95% cases).

We propose to adjust α_j adaptively, so we don't ignore the geometry of polytope $\mathcal{P}$.

Our aim is to minimize variance from (9) over α coming from $q = q_\alpha$. Optimizing the integral $\int \frac{\mathbb{I}(x \in H)p^2(x)}{q(x)} dx$ is equal to the optimization of the variance (9). It is hard to optimize integral so we will optimize its unbiased estimate: $f(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbb{I}(x_i \in H)p^2(x_i)}{q_\alpha(x_i) q_{\alpha'}(x_i)}, \quad x_i \sim q_{\alpha'}$, where α' is a fixed vector.

Let denote $r_\alpha(x) = \frac{q_\alpha(x)}{p(x)} = \sum_{j=1}^{J} \alpha_j H_j(x) P_j^{-1}$.

We have the following optimization task:

$$f(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \frac{H(x_i)}{r_\alpha(x_i) r_{\alpha'}(x_i)} \to \min_\alpha,$$

$$\sum_{j=1}^{J} \alpha_j = 1, \alpha_j \geq \varepsilon. \quad (14)$$

Here conditions $\alpha_j \geq \varepsilon$ are introduced otherwise the condition (10) can be violated.

One can show that $f(\alpha)$ is a convex function on the polyhedral set $S = \{\alpha \big| \sum_{j=1}^{J} \alpha_j = 1, \alpha_j \geq \varepsilon\}$. We propose to apply Frank-Wolfe (FW) method to solve (14) as it is efficient when the set S is a polytope.

Theorem

We define $V(\alpha)$ to be the per-sample variance of the IS estimate of μ with sampling distribution $q = q_\alpha$.

$$V(\alpha) = \text{Var}\left(\frac{\mathbb{I}(x_1 \in H)p(x_1)}{q_\alpha(x_1)}\right) = \quad (15)$$

$$= \int \frac{\mathbb{I}(x \in H)p^2(x)}{q_\alpha(x)} dx - \mu^2, \quad (16)$$

where $x_1 \sim q_\alpha$. We write V^* to denote the optimal over α value of $V(\alpha)$.

Theorem 1

$$\textit{Var}\left(\widehat{\mu}_{q_{\alpha(k)}}\right) \leq \frac{1}{n}\left(V^* + \frac{4L}{k+2}\right),$$

where is L is a Lipschitz constant for $\nabla f(\alpha)$ from (14).

Proof.

1. Firstly, we would like to show the convexity of $f(\alpha)$.

Let's write the Hessian:

$$\frac{\partial^2 f(\alpha)}{\partial\alpha_k \partial\alpha_l} = \frac{1}{n} \sum_{i=1}^{n} \frac{2H_k(x_i)H_l(x_i)P_k^{-1}P_l^{-1}}{r_\alpha^3(x_i) r_{\alpha'}(x_i)}.$$

Our Hessian can be represented as

$$\nabla^2 f(\alpha) = \frac{1}{n} \sum_{i=1}^{n} z_i(\alpha) z_i(\alpha)^{\mathsf{T}},$$

where

$$z_i(\alpha) = \frac{\sqrt{2}}{\sqrt{r_\alpha^3(x_i) r_{\alpha'}(x_i)}} h_i,$$

where $h_i \in \mathbb{R}^J$ and j-th component of vector h_i is $h_i^{(j)} = H_j(x_i)P_j^{-1}$.

Hence our Hessian is the sum of positive-definite matrices, which implies that the Hessian itself is positive semi-definite, consequently $f(\alpha)$ is convex.

2. Now we are going to find Lipschitz constant of $\nabla f(\alpha)$ by showing that

$$\nabla^2 f(\alpha) \preceq L I_n.$$

For this purpose let's find $\lambda_{max}\left(z_i(\alpha) z_i(\alpha)^{\mathsf{T}}\right)$ for some arbitrary $i \in \overline{1,n}$. Actually, $\lambda_{max}\left(z_i(\alpha) z_i(\alpha)^{\mathsf{T}}\right) = z_i(\alpha)^T z_i(\alpha)$ and

$$z_i(\alpha)^T z_i(\alpha) = \frac{2\|h_i\|_2^2}{r_\alpha^3(x_i) r_{\alpha'}(x_i)} \leq$$

$$\leq \frac{2\left(\sum_{j=1}^{J} P_j^{-2}\right)}{r_\alpha^3(x_i) r_{\alpha'}(x_i)} \leq$$

$$\leq \frac{2P_{\max}^4}{\varepsilon^4} \left(\sum_{j=1}^{J} P_j^{-2}\right).$$

The first inequality follows from $\|h_i\|_2^2 = \sum_{j=1}^{J} \left(P_j^{-1} H_j(x_i)\right)^2 \leq \sum_{j=1}^{J} P_j^{-2}$. The second inequality is due to $r_\alpha(x_i) \geq \alpha_j P_j^{-1}$ for some of those

j where $H_j(x_i)$ (at least one such j exists because we generate x_i from the mixture of conditional distributions). That leads to $r_\alpha(x_i) \geq \alpha_j P_j^{-1} \geq \varepsilon P_{\max}^{-1}$ where $P_{\max}$ is the maximum value from P_j, $j \in \overline{1, J}$. Similarly, $r_{\alpha'}(x_i) \geq \alpha'_j P_j^{-1} \geq \varepsilon P_{\max}^{-1}$.

Let L be

$$L = \frac{2P_{\max}^4}{\varepsilon^4} \left(\sum_{j=1}^{J} P_j^{-2} \right).$$

Then

$$\nabla^2 f(\alpha) \preceq \left(\frac{1}{n} \sum_{i=1}^{n} L \right) I_n = L I_n..$$

So we have proved that L is a Lipschitz constant for $\nabla f(\alpha)$.

3. With the help of the Theorem 1 from (Jaggi 2013) we have:

$$f\left(\alpha^{(k)}\right) - f^* \leq \frac{2LD^2}{k+2} \leq \frac{4L}{k+2}, \tag{17}$$

where $D = diam(S) \leq \sqrt{2}$, $S = \{\alpha \big| \sum_{j=1}^{J} \alpha_j = 1, \alpha_j \geq \varepsilon\}$.

4. The last step. We use that $f(\alpha)$ is an unbiased estimate of the integral $\int \frac{\mathbb{I}(x \in H) p^2(x)}{q_\alpha(x)}$:

$$\mathsf{E}_{q_{\alpha'}} f(\alpha^{(k)}) = \int \frac{\mathbb{I}(x \in H) p^2(x)}{q_{\alpha^{(k)}}(x)} dx. \tag{18}$$

Similarly,

$$\mathsf{E}_{q_{\alpha'}} f(\alpha^*) = \int \frac{\mathbb{I}(x \in H) p^2(x)}{q_{\alpha^*}(x)} dx = V^* + \mu^2. \tag{19}$$

$$\begin{aligned}
&\text{Var}\left(\widehat{\mu}_{q_{\alpha^{(k)}}}\right) = \\
&= \frac{1}{n} \left(\int \frac{\mathbb{I}(x \in H) p^2(x)}{q_{\alpha^{(k)}}(x)} dx - \mu^2 \right) = \\
&= \frac{1}{n} \left(\mathsf{E}_{q_{\alpha'}} f(\alpha^{(k)}) - \mu^2 \right) \leq \\
&\leq \frac{1}{n} \left(\mathsf{E}_{q_{\alpha'}} f(\alpha^*) + \frac{4L}{k+2} - \mu^2 \right) = \\
&= \frac{1}{n} \left(V^* + \frac{4L}{k+2} \right).
\end{aligned}$$

Here the first equality is just the definition of variance (9) where $q_{\alpha^{(k)}}$ is used instead of q. The second equality follows from (18). The inequality follows from (17) (mathematical expectation was applied to inequatily). The last eqaulity follows from (19).

Results

We compare the performance of FW (step size policy is $\gamma_k = \frac{2}{k+2}$), SGD ($\gamma_k = \frac{1}{k+1}$) and ALOE (no optimization) methods both on simulated data (pyramid $\mathcal{P}_{\theta,J,\tau}$ where $\theta = 0.01$, $J = 1000$, $\tau = 4$) and real data (Pegase9241 case from MATPOWER (Zimmerman, Murillo-Sanchez, and Thomas 2011), $J = 293$). The number of samples n from formula (14) in all experiments is $n = 10000$. Initial $\alpha^{(0)}$ and α' for FW and SGD: $\alpha^{(0)} = \alpha' = \left(\frac{1}{J}, \ldots, \frac{1}{J}\right)$.

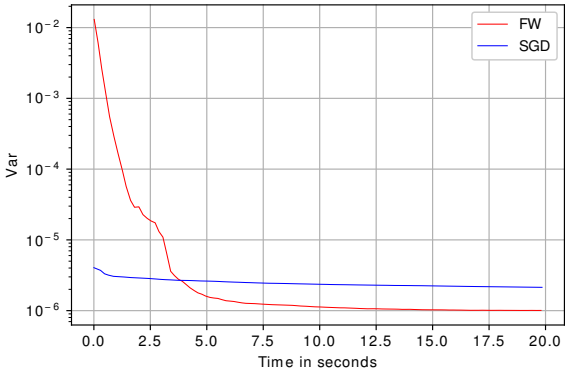

Figure 1: Convergenge based on variance, real data

	FW	SGD	ALOE
simulated data	4.8×10^{-9}	5.3×10^{-8}	3.8×10^{-2}
real data	1.0×10^{-6}	2.1×10^{-6}	1.03×10^{-2}

Table 1: Final variance

The dependence of variance (9) on real time for SGD and FW on real data is shown at the figure 1. The final variance of all 3 methods after optimization for fixed time is shown in the table 1.

We can see from the table 1 that SGD and FW perform better and give less variance than ALOE. That is because FW and SGD make optimization over α so they take into account the geometry of the polytope $\mathcal{P}$. We can see from the figure 1 that FW performs better than SGD because it uses special properties of the set S (S – is a polyhedral set).

References

Jaggi, M. 2013. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In Dasgupta, S., and McAllester, D., eds., *Proceedings of the 30th International Conference on Machine Learning*, number 1 in Proceedings of Machine Learning Research, 427–435. Atlanta, Georgia, USA: PMLR.

Owen, A. B.; Maximov, Y.; and Chertkov, M. 2017. Importance sampling the union of rare events with an application to power systems analysis. *ArXiv e-prints*.

Zimmerman, R. D.; Murillo-Sanchez, C. E.; and Thomas, R. J. 2011. Matpower: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on Power Systems* 26(1):12–19.