arXiv:1710.06965v4 [stat.CO] 19 Dec 2018

Importance sampling the union of rare events with an application to power systems analysis

Art B. Owen
Stanford University

Yury Maximov
T-4 and CNLS, Los Alamos National Laboratory

Michael Chertkov
T-4 and CNLS, Los Alamos National Laboratory

December 2018

Abstract

We consider importance sampling to estimate the probability μ of a union of J rare events H_j defined by a random variable $\boldsymbol{x}$. The sampler we study has been used in spatial statistics, genomics and combinatorics going back at least to Karp and Luby (1983). It works by sampling one event at random, then sampling $\boldsymbol{x}$ conditionally on that event happening and it constructs an unbiased estimate of μ by multiplying an inverse moment of the number of occuring events by the union bound. We prove some variance bounds for this sampler. For a sample size of n, it has a variance no larger than $\mu(\bar{\mu}-\mu)/n$ where $\bar{\mu}$ is the union bound. It also has a coefficient of variation no larger than $\sqrt{(J+J^{-1}-2)/(4n)}$ regardless of the overlap pattern among the J events. Our motivating problem comes from power system reliability, where the phase differences between connected nodes have a joint Gaussian distribution and the J rare events arise from unacceptably large phase differences. In the grid reliability problems even some events defined by 5772 constraints in 326 dimensions, with probability below 10^{-22}, are estimated with a coefficient of variation of about 0.0024 with only $n = 10{,}000$ sample values.

1 Introduction

In this paper we consider a mixture importance sampling strategy to estimate the probability that one or more of a set of rare events takes place. The sampler repeatedly chooses a rare event at random, and then samples the system conditionally on that one event taking place. For each such sample, the total number of occuring events is recorded and a certain reciprocal moment of them is used in the estimate.

This method is a special case of an algorithm in Adler et al. (2008, 2012) for computing exceedance probabilities of Gaussian random fields. It was used earlier by Shi et al. (2007) and Naiman and Priebe (2001) for extrema of genomic scan statistics. Priebe et al. (2001) used it for extrema of some spatial statistic involving marked point processes. The earliest uses that we know are in the computer science literature for enumeration problems like estimating the cardinality of the union of a given list of finite sets. See Karp and Luby (1983) and Frigessi and Vercellis (1985). The above cited papers refer to this method as importance sampling. To distinguish it from other samplers, we will call it ALOE for "At Least One rare Event".

We develop general bounds for the variance of the ALOE importance sampler, and for its coefficient of variation. It has a sampling standard deviation that is no more than some modest multiple of the event probability. This is an especially desirable property in rare event settings. For background on importance sampling of rare events see L'Ecuyer et al. (2009).

Our motivating context is the reliability of the electrical grid when subject to random inputs, such as variable demand by users and variable production, as occurs at wind farms. The rare events describe unacceptably large electrical phase differences at pairs of connected nodes in the grid.

It is common to use a simplified linear direct current (DC) model of the electrical grid, because the equations describing alternating current (AC) are significantly more difficult to work with, and some authors (e.g., Van den Bergh et al. (2014)) find that there is little to be gained from the complexity of an AC model. This DC model is presented in Sauer and Christensen (1984) and Stott et al. (2009). It is also common to model the randomness in the grid as Gaussian, especially over short time horizons.

We make both of these simplifications: linearity and Gaussianity. The probability we consider can then be written

$$\mu = \Pr\bigl(\cup_{j=1}^{J} H_j\bigr), \quad H_j = \{\boldsymbol{x}^{\mathsf{T}}\omega_j \geqslant \tau_j\}, \quad \text{where } \boldsymbol{x} \sim \mathcal{N}(\eta, \Sigma). \tag{1}$$

Section 2 introduces more notation for problem (1) and develops the ALOE sampler as an especially convenient version of mixture importance sampling. In this setting we can compute the union bound $\bar{\mu} = \sum_{j=1}^{J} \Pr(H_j) \geqslant \mu$. Theorem 1 proves that the ALOE estimate $\hat{\mu}$ has variance at most $\mu(\bar{\mu} - \mu)/n$ when n IID samples are used. This can be much smaller than $\mu(1-\mu)/n$ which arises from sampling the nominal distribution of $\boldsymbol{x}$. Section 3 discusses some further sampling properties of our estimator that hold without the Gaussian assumption. When there are J events, the variance of $\hat{\mu}$ is at most $(J + J^{-1} - 2)\mu^2/(4n)$ when the system is sampled n times. Section 4 compares ALOE to a state of the art code `mvtnorm` (Genz et al., 2017) for estimating the probability that a multivariate Gaussian of up to 1000 variables with arbitrary covariance belongs to a given hyperrectangle. ALOE is simpler and extends to higher dimensions. When we studied rare event cases, ALOE was more accurate. In our examples that are not rare events, `mvtnorm` was more accurate. We also make a comparison to a directional sampling method studied recently by Ahn and Kim (2018). That

method is far better than ALOE on our low dimensional test problems but very seriously underestimates the rare event probability on our high dimensional test problems. Section 5 describes the power system application. Section 6 contains some discussions. The appendix proves Theorem 1 for any set of J events, not just those given by a Gaussian distribution. The theorem applies so long as we can sample conditionally on any one event H_j and then determine which other events H_ℓ also occur. We finish this section with some comments and some references.

One common way for rare event sampling to be inaccurate is that we might fail to obtain any points where the rare event happens. That leads to a severe under-estimation of the rare event probability. In ALOE, the corresponding problem is the failure to sample any points where two or more of the rare constituent events occur. In that case ALOE will return the union bound as the estimated rare event probability instead of zero. That is also a setting where the union bound is likely to be a good approximation. So ALOE is robust against severe underestimates of the rare event probability. The second common problem for rare event sampling is an extreme value of the likelihood ratio weighting applied to the observations. In ALOE, the largest possible weight is only J times as large as the smallest one.

Our sampler is closely related to instanton methods in power systems engineering. See Chertkov, Pan et al. (2011), Chertkov, Stepanov et al. (2011), and Kersulis et al. (2015). Out of all the configurations of random inputs to a system, the most probable one causing the failure is called the instanton. When there are thousands of failure types there are correspondingly thousands of instantons, each one a conditional mode of the distribution of $\boldsymbol{x}$. Our initial thought was to do importance sampling from a mixture of distributions, with each mixture component defined by shifting the Gaussian distribution's mean to an instanton. By sampling conditionally on an event, ALOE avoids wasting samples outside the failure region. By conditioning instead of shifting, we get better control over the likelihood ratio in the importance sampler.

ALOE is a form of multiple importance sampling. Multiple importance sampling originated in computer graphics (Lafortune and Willems, 1993; Veach and Guibas, 1994). Owen and Zhou (2000) found a useful way to combine it with control variates defined by the mixture components. Elvira et al. (2015a,b) investigate computational efficiency of some mixture importance sampling and weighting strategies.

We do not consider self-normalized importance sampling (SNIS) in this paper. SNIS is useful in settings where we can compute an unnormalized version of our target density but cannot sample from it efficiently, if at all. SNIS is common in Bayesian applications (Liu, 2001, Chapter 2). For a recent adaptive version of SNIS, see Cornuet et al. (2012). For rare event estimation, we show in the Appendix that self-normalized importance sampling cannot deliver a coefficient of variation meaningfully below $2/\sqrt{n}$ asymptotically. The optimal sampler for SNIS allocates precisely half of its probability in the rare event and half outside of it. The optimal plain IS estimator, by contrast, places all of its probability on the rare event and has zero variance. Ordinary importance

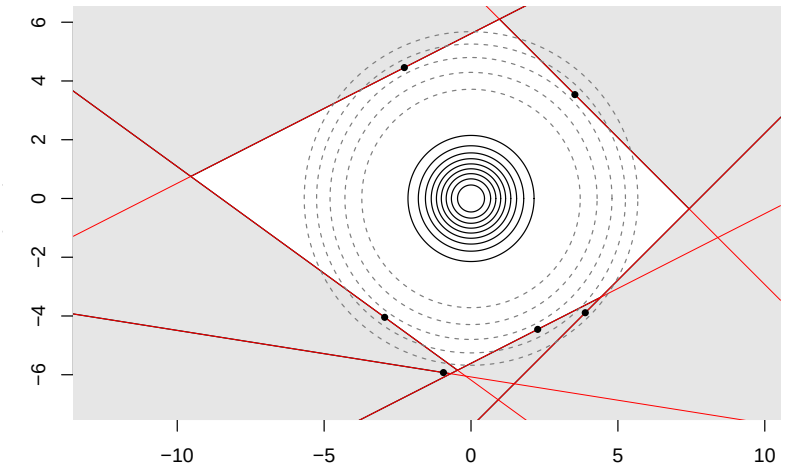


Figure 1: The solid circles contain 10%, 20% up to 90% of the $\mathcal{N}(0, I)$ distribution. The dashed circles contain all but 10^{-k} of the probability for $3 \leqslant k \leqslant 7$. The six solid lines denote half-spaces. The solid points are the corresponding conditional modes (instantons). The rare event of interest is $\boldsymbol{x}$ in the shaded region, when $\boldsymbol{x} \sim \mathcal{N}(0, I)$.

sampling can attain much smaller variances, and so we focus on it for the rare event problem.

2 Gaussian case

For concreteness, we present ALOE first for Gaussian random variables. The earliest use we have seen for Gaussian variables is Adler et al. (2008). We let $\boldsymbol{x} \in \mathbb{R}^d$ have the standard Gaussian distribution, $\mathcal{N}(0, I)$, deferring general Gaussians to Section 2.1. We are interested in computing the probability that $\boldsymbol{x}$ lies outside a polytope $\mathcal{P}$. In our motivating applications, the interior of the polytope defines a safe operating region and we assume that $\boldsymbol{x} \notin \mathcal{P}$ is a rare event. For $j = 1, \ldots, J$, define half-spaces

$$H_j = \{\boldsymbol{x} \mid \omega_j^{\mathsf{T}} \boldsymbol{x} \geqslant \tau_j\}$$

where each $\tau_j \in \mathbb{R}$ and $\omega_j \in \mathbb{R}^d$, with $\omega_j^{\mathsf{T}} \omega_j = 1$. Then $\mathcal{P} = \cap_{j=1}^J H_j^c$ and we want to find $\mu = \Pr(\boldsymbol{x} \in H)$ where $H = \cup_{j=1}^J H_j = \mathcal{P}^c$. The set $\mathcal{P}$ is convex and not necessarily bounded. Ordinarily $\tau_j > 0$, because we are interested in rare events.

The setting is illustrated in Figure 1 for $J = 6$ half-spaces. In that example, two of the half-spaces have their conditional modes inside the union of the other

half-spaces. One of those half-spaces is entirely included in the union of the others.

Letting $P_j = \Pr(\boldsymbol{x} \in H_j) = \Phi(-\tau_j)$, we know that

$$\max_{1 \leqslant j \leqslant J} P_j =: \underline{\mu} \leqslant \mu \leqslant \bar{\mu} := \sum_{j=1}^{J} P_j. \tag{2}$$

The right hand side is the union bound which is sometimes very conservative and sometimes quite accurate.

We will need to use some inclusion-exclusion formulas, so some notation for these follows. For any $u \subseteq 1{:}J \equiv \{1, 2, \ldots, J\}$, let $H_u = \cup_{j \in u} H_j$, so $H_j = H_{\{j\}}$ and by convention $H_\varnothing = \varnothing$. We identify the set H_u with the function $H_u(\boldsymbol{x}) = \mathbf{1}\{\boldsymbol{x} \in H_u\}$. Next define $P_u = \mathbb{E}(H_u(\boldsymbol{x}))$ for $\boldsymbol{x} \sim \mathcal{N}(0, I)$. We use $-u$ for complements in $1{:}J$ especially within subscripts, and $H_u^c(\boldsymbol{x})$ for the complementary outcome $1 - H_u(\boldsymbol{x})$. Let $S(\boldsymbol{x}) = \sum_{j=1}^{J} H_j(\boldsymbol{x})$ count the number of rare events that happen. For $s = 0, 1, \ldots, J$, let $T_s = \Pr(S = s)$ give the distribution of S. We use $|u|$ for the cardinality of u. Our estimand is

$$\mu = P_{1:J} = \sum_{|u|>0} (-1)^{|u|-1} P_u, \tag{3}$$

by inclusion-exclusion.

Here we motivate ALOE as an especially simple mixture sampler. The mixture components we use are conditional distributions $q_j = \mathcal{L}(\boldsymbol{x} \mid \omega_j^{\mathsf{T}} \boldsymbol{x} \geqslant \tau_j)$, for $j = 1, \ldots, J$. They have probability density functions $q_j(\boldsymbol{x}) = p(\boldsymbol{x}) H_j(\boldsymbol{x}) / P_j$.

Let $\alpha_1, \ldots, \alpha_J$ be nonnegative numbers summing to 1, and $q_\alpha = \sum_{j=1}^{J} \alpha_j q_j$. A mixture importance sampling estimate of μ based on n draws $\boldsymbol{x}_i \sim q_\alpha$ is

$$\hat{\mu}_\alpha = \frac{1}{n} \sum_{i=1}^{n} \frac{p(\boldsymbol{x}_i) H_{1:J}(\boldsymbol{x}_i)}{\sum_{j=1}^{J} \alpha_j q_j(\boldsymbol{x}_i)} = \frac{1}{n} \sum_{i=1}^{n} \frac{H_{1:J}(\boldsymbol{x}_i)}{\sum_{j=1}^{J} \alpha_j H_j(\boldsymbol{x}_i) P_j^{-1}}. \tag{4}$$

Notice that $p(\boldsymbol{x}_i)$ has conveniently canceled from numerator and denominator. Although the inclusion-exclusion formula (3) contains $2^J - 1$ nonzero terms, each summand in the unbiased estimate in (4) can be computed at cost $O(J)$.

We can induce further cancellation in (4) by making α_j / P_j constant in j. Taking $\alpha_j = \alpha_j^* \equiv P_j / \bar{\mu}$, we get

$$\hat{\mu}_{\alpha^*} = \frac{\bar{\mu}}{n} \sum_{i=1}^{n} \frac{H_{1:J}(\boldsymbol{x}_i)}{\sum_{j=1}^{J} H_j(\boldsymbol{x}_i)} = \frac{\bar{\mu}}{n} \sum_{i=1}^{n} \frac{1}{S(\boldsymbol{x}_i)}, \quad \boldsymbol{x}_i \overset{\text{iid}}{\sim} q_{\alpha^*}, \tag{5}$$

because $H_{1:J}(\boldsymbol{x}) = 1$ always holds for $\boldsymbol{x} \sim q_{\alpha^*}$. The estimate (5) is a multiplicative adjustment to the union bound $\bar{\mu}$. The terms $S(\boldsymbol{x}_i)^{-1}$ range from 1 to $1/J$ and so we will never get $\hat{\mu}_{\alpha^*}$ larger than the union bound or smaller than $\bar{\mu}/J$. This is convenient because $\bar{\mu} \geqslant \mu \geqslant \bar{\mu}/J$ always holds.

Theorem 1. *Let $\hat{\mu}_{\alpha^*}$ be given by* (5)*. Then*

$$\mathbb{E}(\hat{\mu}_{\alpha^*}) = \mu, \tag{6}$$

and

$$\mathrm{Var}(\hat{\mu}_{\alpha^*}) = \frac{1}{n}\biggl(\bar{\mu}\sum_{s=1}^{J}\frac{T_s}{s} - \mu^2\biggr) \leqslant \frac{\mu(\bar{\mu}-\mu)}{n}. \tag{7}$$

Proof. See the appendix, where this is proved for a general set of J events, not necessarily from Gaussian half-spaces. □

The upper bound (7) involves the unknown μ, so it is not available for planning purpose when we want to select n. The variance and the coefficient of variation, $\mathrm{cv}(\hat{\mu}_{\alpha^*}) = \mathrm{Var}(\hat{\mu}_{\alpha^*})^{1/2}/\mu$ can both be bounded in terms of known quantities $\bar{\mu}$ and $\underline{\mu}$ from (2) as follows.

Corollary 1. *Let $\hat{\mu}_{\alpha^*}$ be given by* (5)*. Then* $\mathrm{Var}(\hat{\mu}_{\alpha^*}) \leqslant \bar{\mu}^2/(4n)$*. If $\underline{\mu} \geqslant \bar{\mu}/2$ then also* $\mathrm{Var}(\hat{\mu}_{\alpha^*}) \leqslant \underline{\mu}(\bar{\mu}-\underline{\mu})/n$*. Similarly,*

$$\mathrm{cv}(\hat{\mu}_{\alpha^*}) \leqslant \frac{1}{\sqrt{n}}\min\Bigl\{\sqrt{\bar{\mu}/\underline{\mu}-1}, \sqrt{J-1}\,\Bigr\}. \tag{8}$$

Proof. The claims about $\mathrm{Var}(\hat{\mu}_{\alpha^*})$ follow from maximizing (7) over $\mu \in [\,\underline{\mu}, \bar{\mu}]$. Next $\mathrm{cv}(\hat{\mu}_{\alpha^*})^2 = (\bar{\mu}-\mu)/(n\mu) = (\bar{\mu}/\mu - 1)/n$. Then (8) follows because $\mu \geqslant \underline{\mu}$ and $\mu \geqslant \bar{\mu}/J$. □

A rare event estimator has bounded relative error if $\mathrm{cv}(\hat{\mu})$ remains bounded as one takes the limit in a sequence of problems (Asmussen and Glynn, 2007, Chapter VI). The sequence is typically one where the event of interest becomes increasingly rare, for instance as $\mu \to 0$ in the present context. Corollary 1 provides a bounded relative error property for ALOE in that limit or indeed in any sequence of problems where J/n is uniformly bounded.

If the product $H_u(\boldsymbol{x})H^c_{-u}(\boldsymbol{x})$ equals one then it means that $\boldsymbol{x} \in H_j$ if and only if $j \in u$. We use this to write the union bound in a useful way:

$$\bar{\mu} = \sum_{j=1}^{J}\Pr(H_j(\boldsymbol{x})) = \sum_{j=1}^{J}\sum_{u\subseteq 1:J}\mathbb{E}\bigl(H_u(\boldsymbol{x})H^c_{-u}(\boldsymbol{x})\bigr)1_{j\in u} = \sum_{s=1}^{J} sT_s.$$

That is $\bar{\mu} = \mathbb{E}(S(\boldsymbol{x})) = \mu\mathbb{E}(S(\boldsymbol{x}) \mid S(\boldsymbol{x}) > 0)$ and so we may write (7) as

$$\mathrm{Var}(\hat{\mu}_{\alpha^*}) = \frac{\mu^2}{n}\Bigl(\mathbb{E}(S \mid S>0)\mathbb{E}(S^{-1} \mid S>0) - 1\Bigr). \tag{9}$$

We will use (9) in Section 3 to get additional bounds.

2.1 General Gaussians

Now suppose that we are given $\boldsymbol{y} \sim \mathcal{N}(\eta, \Sigma)$ and the half-spaces are defined by $\gamma_j^{\mathsf{T}}\boldsymbol{y} \geqslant \kappa_j$. We assume that Σ is nonsingular. If it is not, then we can reduce $\boldsymbol{y}$ to a subset of components whose variance is nonsingular, and write the other components as linear functions of this reduced set. We also assume that we can afford to take a matrix square root $\Sigma^{1/2}$. Now $\boldsymbol{x} = \Sigma^{-1/2}(\boldsymbol{y} - \eta) \sim \mathcal{N}(0, I)$, and $\boldsymbol{y} = \eta + \Sigma^{1/2}\boldsymbol{x}$. Then the half-spaces are given by

$$\omega_j^{\mathsf{T}}\boldsymbol{x} \geqslant \tau_j, \quad \text{where} \quad \omega_j = \frac{\gamma_j^{\mathsf{T}}\Sigma^{1/2}}{\sqrt{\gamma_j^{\mathsf{T}}\Sigma\gamma_j}}, \quad \text{and} \quad \tau_j = \frac{\kappa_j - \gamma_j^{\mathsf{T}}\eta}{\sqrt{\gamma_j^{\mathsf{T}}\Sigma\gamma_j}},$$

for $\boldsymbol{x} \sim \mathcal{N}(0, I)$. For rare events, we will have $\kappa_j > \gamma_j^{\mathsf{T}}\eta$. In some of our motivating contexts one must optimize a cost over η. Here we remark that changes to η change τ_j but not ω_j.

2.2 Sampling algorithms

We want to sample $\boldsymbol{x} \sim \mathcal{N}(0, I)$ conditionally on $\boldsymbol{x}^{\mathsf{T}}\omega \geqslant \tau$ for a unit vector ω and scalar τ. We can use the following steps:

1) Sample $\boldsymbol{z} \sim \mathcal{N}(0, I)$.
2) Sample $u \sim \mathbb{U}(0, 1)$.
3) Let $y = \Phi^{-1}(\Phi(\tau) + u(1 - \Phi(\tau)))$.
4) Deliver $\boldsymbol{x} = \omega y + (I - \omega\omega^{\mathsf{T}})\boldsymbol{z}$.

These steps can be justified by the analysis in Doucet (2010) who attributes the algorithm to the astrophysics literature. Step 3 replaces a $\mathcal{N}(0, 1)$ distribution for $y = \boldsymbol{x}^{\mathsf{T}}\omega$ by a truncated Gaussian random variable obtained via inversion.

The algorithm above can be problematic numerically when $\Phi(\tau)$ is close to 1 as it will be for very rare events. For instance, in the R language (R Core Team, 2015), $\Phi(10)$ yields 1 and then $\Phi^{-1}(\Phi(10) + u(1 - \Phi(10)))$ yields ∞ for any u. Some of our electrical grid examples have $\max_j \tau_j > 10^{10}$. That is, some of the potential failure modes are virtually impossible.

Because $\tau > 0$ might be quite large, we get better numerical stability by sampling $\boldsymbol{x} \sim \mathcal{N}(0, I)$ conditionally on $\boldsymbol{x}^{\mathsf{T}}\omega \leqslant -\tau$ and then delivering $-\boldsymbol{x}$. The advantage of simulating extreme Gaussians this way was goes back at least to Cunningham (1969) and may well be older than that. The steps are as follows:

1) Sample $\boldsymbol{z} \sim \mathcal{N}(0, I)$.
2) Sample $u \sim \mathbb{U}(0, 1)$.
3) Let $y = \Phi^{-1}(u\Phi(-\tau))$.
4) Let $\boldsymbol{x} = \omega y + (I - \omega\omega^{\mathsf{T}})\boldsymbol{z}$.
5) Deliver $\boldsymbol{x} = -\boldsymbol{x}$.

Even a very small $u = 10^{-12}$ combined with $\tau = 10$ yields

$$\Phi^{-1}(10^{-12} \times \Phi(-10)) \doteq \Phi^{-1}(7.62 \times 10^{-36}) \doteq -12.44$$

without any underflow in the R language (R Core Team, 2015). In cases with extremely large τ_j we will ordinarily get $P_j = 0$ and then never sample condi-

tionally on the corresponding H_j. We compute step 4 via $\boldsymbol{x} = \omega y + z - \omega(\omega^{\mathsf{T}}\boldsymbol{z})$ to avoid a potentially expensive multiplication $(I - \omega\omega^{\mathsf{T}})\boldsymbol{z}$.

3 Importance sampling properties

As shown in the Appendix, Theorem 1 holds more generally than the Gaussian case. In this more general setting, we have J events, H_j, on a common sample space $\mathcal{X}$ where $\boldsymbol{x} \in \mathcal{X}$ has probability density p. Event H_j has probability P_j. As before, we want $\mu = \Pr(H)$ where $H = \cup_j H_j$ and the union bound is $\mu \leqslant \bar{\mu} = \sum_j P_j$. We assume that $0 < \bar{\mu} < \infty$. The upper bound only has to be checked if $J = \infty$. If $\bar{\mu} = 0$, then we know $\mu = 0$ without any sampling.

When we sample, we ensure that at least one rare event takes place every time, by first picking an event H_j with probability proportional to P_j. Then we sample $\boldsymbol{x} \in \mathcal{X}$ conditionally on H_j and find $S(\boldsymbol{x}) = \sum_{\ell=1}^{J} H_\ell(\boldsymbol{x})$, the total number of events that occur. This includes H_j and so our sample values always have $S(\boldsymbol{x}_i) \geqslant 1$. The importance sampling estimate $\hat{\mu}_{\alpha^*}$ averages $\bar{\mu}/S(\boldsymbol{x}_i)$ over n independent replicates. As in the prior section, we use

$$T_s = \Pr(S(\boldsymbol{x}) = s) = \int_{\mathbb{R}^d} 1\{S(\boldsymbol{x}) = s\} p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$

for the probability of exactly s events happening. Then the variance of $\hat{\mu}$ is given by (7).

The optimal importance sampling distribution for estimating μ is uniform on $H = \{\boldsymbol{x} \mid H(\boldsymbol{x}) = 1\}$. Sampling from this distribution would yield an estimate with variance zero. Not surprisingly, we are seldom able to do that in applications. The ALOE sampler takes $\boldsymbol{x} \in \mathcal{X}$ with probability proportional to $S(\boldsymbol{x})$, so it has support set H.

We think that many applications will have events H_j that rarely co-occur. In that case $S(\boldsymbol{x})$ is nearly constant at 1 for $\boldsymbol{x} \in H$, and the ALOE sampler is close to the optimal importance sampler. Other applications may have a few near duplicated events that co-occur often. One extreme setting has a common cause that triggers all J events at once and those events almost never arise outside of that common situation. In that case $S(\boldsymbol{x})$ is again nearly constant on H, this time usually equal to J, and ALOE is again nearly optimal.

The variance bound $\mu(\bar{\mu} - \mu)/n$ from (7) can be conservative. It stems from $T_s/s \leqslant T_s$, when $s \geqslant 1$. If $\Pr_{\alpha^*}(S > 1)$ is appreciably large then the variance can be meaningfully less than that bound. We can improve the variance bound by using the following lemma.

Lemma 1. *Let S be a random variable supported on $\{1, 2, \ldots, J\}$ for $J \in \mathbb{N}$. Then*

$$\mathbb{E}(S)\mathbb{E}(S^{-1}) \leqslant \frac{J + J^{-1} + 2}{4} \tag{10}$$

with equality if and only if $S \sim \mathbb{U}\{1, J\}$.

Proof. See the appendix. □

Lemma 1 tells us that for $J \geqslant 2$, our worst case setting is one where half of the time that one or more events happen, exactly one happens and half of the time, all J of them happen. While that is not plausible for Gaussian $\boldsymbol{x}$ and large J it can indeed happen for combinatorial enumeration problems like those of Karp and Luby (1983). From Theorem 2 and Lemma 1, we get

$$\mathrm{Var}(\hat{\mu}_{\alpha^*}) = \frac{\mu^2}{n}\left(\left(\sum_{s=1}^{J} s\frac{T_s}{\mu}\right)\left(\sum_{s=1}^{J} s^{-1}\frac{T_s}{\mu}\right) - 1\right) \leqslant \frac{\mu^2}{n}\frac{J + J^{-1} - 2}{4}, \tag{11}$$

because T_s/μ is a probability distribution on $\{1, 2, \ldots, J\}$.

Sometimes we are interested in the probability of sub-events of H. Let $f(\boldsymbol{x})$ be supported on H and define $\nu(f) = \nu = \int f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_H f(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. We may use ALOE, via

$$\hat{\nu} = \frac{\bar{\mu}}{n}\sum_{i=1}^{n}\frac{f(\boldsymbol{x}_i)}{S(\boldsymbol{x}_i)}, \quad \boldsymbol{x}_i \overset{\text{iid}}{\sim} q_{\alpha^*}.$$

Then by the same arguments used in the Appendix,

$$\mathbb{E}(\hat{\nu}) = \nu \quad \text{and} \quad \mathrm{Var}(\hat{\nu}) = \frac{1}{n}\left(\bar{\mu}\int_H \frac{f(\boldsymbol{x})^2 p(\boldsymbol{x})}{S(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} - \nu^2\right).$$

If $f(\boldsymbol{x}) \in \{0, 1\}$, then $\mathrm{Var}(\hat{\nu}) \leqslant \nu(\bar{\mu} - \nu)/n$. That is, when f describes a rare event that can only occur if one or more of the H_j also occur, we can reduce its Monte Carlo variance from $\nu(1-\nu)/n$ to at most $\nu(\bar{\mu} - \nu)/n$, in cases where $\bar{\mu} < 1$.

4 Comparisons

Here we consider some numerical examples comparing ALOE to `pmvnorm` from the R package `mvtnorm` (Genz et al., 2017). This package can make use of special properties of the Gaussian distribution, and it works in high dimensions.

We begin by describing `mvtnorm` based on Genz and Bretz (2009) and a personal communication from Alan Genz. The program computes

$$\Pr(\boldsymbol{a} \leqslant \boldsymbol{y} \leqslant \boldsymbol{b}) \equiv \Pr(a_j \leqslant y_j \leqslant b_j,\ j = 1, \ldots, d)$$

for $\boldsymbol{y} \sim \mathcal{N}(\eta, \Sigma)$, where $-\infty \leqslant a_j \leqslant b_j \leqslant \infty$ for $j = 1, \ldots, d$, and Σ can be rank deficient. We can use it to compute $\mu = \Pr(\sum_{j=1}^{J} \mathbf{1}\{\omega_j^{\mathsf{T}}\boldsymbol{x} \geqslant \tau_j\} > 0)$ for $\boldsymbol{x} \sim \mathcal{N}(0, I)$ via

$$1 - \mu = \Pr(\Omega^{\mathsf{T}}\boldsymbol{x} \leqslant \mathcal{T}) = \Pr(\boldsymbol{y} \leqslant \mathcal{T}), \quad \boldsymbol{y} \sim \mathcal{N}(0, \Omega^{\mathsf{T}}\Omega).$$

The code can handle dimensions up to 1000. In our context, that means at most $J = 1000$ half-spaces. The dimension d can be higher. The related `pmvt`

function handles multivariate t random variables. The code has three different algorithms in it. One from Genz (2004) handles two and three dimensional semi-infinite regions, one from Miwa et al. (2003) is for dimensions up to 20 and the rest are handled by an algorithm from Genz and Bretz (2009). This latter algorithm uses a number of methods. It uses randomized Korobov lattice rules as described by Cranley and Patterson (1976) for the first 100 dimensions, in conjunction with antithetic sampling. There are usually 8 randomizations. For more than 100 dimensions it applies a method from Niederreiter (1972). There are a series of increasing sample sizes in use, and the method provides an estimated error (3.5 standard errors) based on the randomization. The approach is via sequential conditional sampling, after strategically ordering the variables (e.g., putting unconstrained ones first). The R package calls a FORTRAN program for the computation, so it is very fast. We use the default implementation which uses up to 25,000 quadrature points.

The main finding in comparison to Genz and Bretz (2009) is that importance sampling is more effective when the polytope of interest is the complement of a rare event. This is not meant to be a criticism of `pmvnorm`. That code was not specifically designed to compute the complement of a rare event. The comparison is relevant because we are not aware of alternative code tuned for the high dimensional rare event cases that we need, and `pmvnorm` is a well regarded and widely available general solution, that seemed to us like the best off-the-shelf tool.

Botev and L'Ecuyer (2015) provide a competing method to Genz and Bretz (2009) for estimating polytope probabilities with Gaussian and t-distributed data. Like Genz and Bretz (2009) they address problems where $\boldsymbol{x} \in \mathcal{P}$ is the rare event, not $\boldsymbol{x} \notin \mathcal{P}$. We have not compared their method numerically to ALOE but we expect that like Genz and Bretz (2009), it will dominate ALOE when the event is not rare but not when the event is rare.

A reviewer asked us to compare our method to the recent work in Ahn and Kim (2018) on computing expectations over a union of half-spaces. Their approach is to pick a unit vector δ uniformly at random in $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d \mid \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} = 1\}$ and average $\Pr(\boldsymbol{x} \in H)$ over the line $\{y\delta \mid y \in \mathbb{R}\}$. This line extends in the positive y reaching the set H at distance $\min_j \tau_j/\omega_j^{\mathsf{T}}\delta_i$ taking the minimum over j with $\omega^{\mathsf{T}}\delta > 0$. Should that set of j be empty, the line never reaches H in the positive y direction. It has a similarly defined extent in the negative y direction. The scale y is a symmetric random variable with $y^2 \sim \chi^2_{(d)}$ because $\|\boldsymbol{x}\|^2 \sim \chi^2_{(d)}$. Putting these together we find that their directional simulation estimator is

$$\hat{\mu}_{\mathrm{AK}} = \frac{1}{2n}\sum_{i=1}^{n} \max_{1\leqslant j\leqslant J} \bar{G}_d\Big(\frac{\tau_j^2}{(\omega^{\mathsf{T}}\delta_i)^2}\Big)\mathbf{1}\{\omega^{\mathsf{T}}\delta_i > 0\} + \max_{1\leqslant j\leqslant J} \bar{G}_d\Big(\frac{\tau_j^2}{(\omega^{\mathsf{T}}\delta_i)^2}\Big)\mathbf{1}\{\omega^{\mathsf{T}}\delta_i < 0\} \tag{12}$$

where $\bar{G}_d$ is one minus the cumulative distribution function of $\chi^2_{(d)}$. They use their estimator on some ellipsoidally symmetric distributions generalizing the Gaussian. We ran directional sampling on the two examples described next,

τ	μ	$\mathbb{E}((\hat{\mu}_{\mathrm{ALOE}}/\mu-1)^2)$	$\mathbb{E}((\hat{\mu}_{\mathrm{MVN}}/\mu-1)^2)$
2	1.35×10^{-01}	0.000399	9.42×10^{-08}
3	1.11×10^{-02}	0.000451	9.24×10^{-07}
4	3.35×10^{-04}	0.000549	2.37×10^{-02}
5	3.73×10^{-06}	0.000600	$1.81\times10^{+00}$
6	1.52×10^{-08}	0.000543	4.39×10^{-01}
7	2.29×10^{-11}	0.000559	3.62×10^{-01}
8	1.27×10^{-14}	0.000540	1.34×10^{-01}

Table 1: Results from 100 computations of $\Pr(\boldsymbol{x} \notin \mathcal{P}(360,\tau))$ for various τ. The true mean μ is very nearly $\exp(-\tau^2/2)$. Importance sampling is more accurate for large τ (rare events), while `pmvnorm` is more accurate for small τ.

and it was extremely good on one of them and extremely inaccurate on the other.

4.1 Circumscribed polygon

Let $\mathcal{P}(J,\tau) \subset \mathbb{R}^2$ be the regular polygon of $J \geqslant 3$ sides circumscribed around the circle of radius $\tau > 0$. This polygon is the intersection of H_j^c where $H_j = \{\boldsymbol{x} \in \mathbb{R}^2 \mid \omega_j^{\mathsf{T}}\boldsymbol{x} \geqslant \tau\}$ where $\omega_j^{\mathsf{T}} = (\sin(2\pi j/J), \cos(2\pi j/J))$, for $j = 1,\dots,J$. We want $\mu = \Pr(\boldsymbol{x} \in \mathcal{P}^c)$ for $\boldsymbol{x} \sim \mathcal{N}(0, I)$. Here we know that $\mu \leqslant \Pr(\chi^2_{(2)} \geqslant \tau^2) = \exp(-\tau^2/2)$. Also, the gap between the circle of radius τ and the circumscribed polygon has area $G(J,\tau) = (J\tan(\pi/J) - \pi)\tau^2$. The bivariate Gaussian density in this gap is at most $\exp(-\tau^2/2)/(2\pi)$. Therefore

$$\Pr(\boldsymbol{x} \in \mathcal{P}^c) \geqslant \exp(-\tau^2/2) - G(J,\tau)\exp(-\tau^2/2)/(2\pi)$$

that is

$$1 \geqslant \frac{\Pr(\boldsymbol{x} \in \mathcal{P}^c)}{\exp(-\tau^2/2)} \geqslant 1 - \frac{G(J,1)\tau^2}{2\pi} \doteq 1 - \frac{\pi^2\tau^2}{6J^2},$$

for large J.

For $J = 360$ and $\tau = 6$, we have $\mu \leqslant \exp(-18) \doteq 1.52 \times 10^{-8}$. The lower bound is about 0.9995 times the upper bound, so we treat the upper bound as exact. Figure 2 shows histograms of 100 simulations of $\hat{\mu}/\mu$ using ALOE and using pvnorm. In this case ALOE is much more accurate. The mean square relative error $\mathbb{E}((\hat{\mu}/\mu-1)^2)$ is about 800-fold smaller for ALOE than pvnorm. We also see that pvnorm has high positive skewness and the histogram of estimates has most of its mass well below the mean.

Table 1 shows summary results for this problem with different values of τ. We see that pvnorm is superior when the event is not rare but ALOE is superior for rare events. The large error for pvnorm with $\tau = 5$ stemmed from a small number of outliers among the 100 trials.

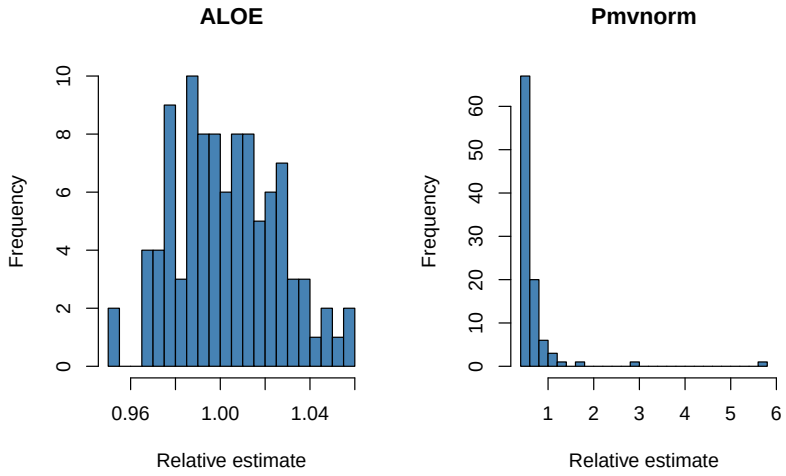


Figure 2: Results of 100 estimates of the $\Pr(\boldsymbol{x} \notin \mathcal{P}(360, 6))$, divided by $\exp(-6^2/2)$. Left panel: ALOE. Right panel: pmvnorm.

The upper bound in equation (7) is $\mathrm{Var}(\hat{\mu}) \leqslant \mu(\bar{\mu} - \mu)/n$, from which $\mathbb{E}((\hat{\mu}/\mu - 1)^2) \leqslant (\bar{\mu}/\mu - 1)/n$. For $\tau = 6$ this yields about 0.022, which is over 20 times the actual mean squared relative error from Table 1.

It is possible that this example is artificially easy for importance sampling, due to the symmetry. Whichever half-space H_j we sample, the distribution of overlapping half-spaces H_k for $k \neq j$ is the same. Two half-spaces differ from H_j by a one degree angle, two differ by a two degree angle and so on. To get a more varied range of overlap patterns, we replaced angles $2\pi j/360$ by angles $2\pi \times p(j)/360$ where $p(j)$ is the j'th prime among integers up to 360. There are 72 of them, of which the largest is 359. With $\tau = 6$ and 100 replications using $n = 1000$ points in importance sampling, we have variance of $\hat{p}/\exp(-18)$ equal to 0.00077. The comparable figure for `mvtnorm` is 8.5. There were a few outliers there including one that was more than 6 times the union bound. The gap between the prime angle polygon and the inscribed circle is larger than the one formed by the full polygon. Pooling all the importance sampling runs leaves an estimate of about $0.94 \times \exp(-18)$ for μ. In this example, we see importance sampling working quite well without symmetry.

The estimator $\hat{\mu}_{\mathrm{AK}}$ is much better than both ALOE and `pmvnorm` for this problem. With only a sample of $n = 100$ it reached essentially double precision accuracy, with a standard error of about 5×10^{-16} on the symmetric polygon. For the problem using prime number angles the standard error was about 3×10^{-12}. This problem is even more artificially easy for directional simulation method than the polygon is for ALOE. The distance from the origin to H is nearly constant over all angles.

4.2 High dimensional half-spaces

The previous example was low dimensional and each of the half-spaces sampled had numerous similar ones, differing in angle by a small number of degrees. Thus μ was quite a bit smaller than $\bar{\mu}$. Here we consider a high dimensional setting where the half-spaces have less overlap.

Two uniform random unit vectors ω_1 and ω_2 in $\mathbb{R}^d$ are very likely to be nearly orthogonal for large d. Then $\boldsymbol{x}^{\mathsf{T}}\omega_j > \tau_j$ are nearly independent events. For independent events, we would have

$$\Pr(\boldsymbol{x} \notin \mathcal{P}) = 1 - \prod_{j=1}^{J}(1 - P_j).$$

To make $\boldsymbol{x} \notin \mathcal{P}$ a rare event, the P_j must be small and then the probability above will be close to the union bound. Theorem 2 predicts good performance for importance sampling here.

For this test 200 sample problems were constructed. The dimensions were chosen by $d \sim \mathbb{U}\{20, 50, 100, 200, 500\}$. Then there were $J \sim \mathbb{U}\{d/2, d, 2d\}$ constraints chosen with uniform random unit vectors $\omega_j \in \mathbb{R}^d$. The threshold τ was chosen so that $\log_{10}$ of the union bound was $\mathbb{U}[4, 8]$, followed by rounding to two significant figures. Then $\hat{\mu}$ was computed by importance sampling with $n = 1000$ samples, and by `pmvnorm`. Figure 3 shows the results. The ALOE sampling value was always very close to the union bound which in turn is essentially equal to what one would see for independent events. The values from `pmvnorm` were usually too small but sometimes far too large, orders of magnitude larger than the union bound. By construction the intersection probabilities are quite rare. In importance sampling, 77.5% of the simulations had no intersections among 1000 trials and the others had only a few intersections. Therefore it is clear that the probabilities should be close to the union bounds here.

We also ran the directional method on these high-dimensional half space problems. It did very poorly because the set H only comes close to the origin at a tiny proportion of the unit vectors $\delta \in \mathbb{S}^{d-1}$. The estimate $\hat{\mu}_{\mathrm{AK}}$ was usually smaller than the known lower bound $\underline{\mu} = \max_{1 \leqslant j \leqslant J} \Phi(-\tau_j)$, sometimes much smaller, in one instance below $10^{-30}\underline{\mu}$. Two hundred results are presented in Figure 4. That estimator was also larger than the known upper bound $\bar{\mu}$ by as much as 40-fold in some simulation. It was less severe in that regard than `pmvnorm`. Because it performed so poorly we did not implement it on the power systems problem in the next section.

5 Power system infeasibility

5.1 Model

Our power system models are based on a network of N nodes, called busses. Some busses put power into the network and others consume power. The M

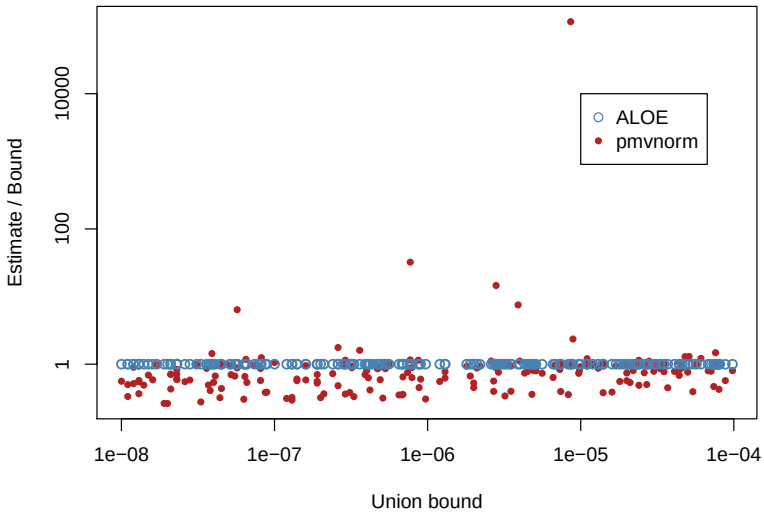


Figure 3: Results of 200 estimates of the μ for varying high dimensional problems with nearly independent events.

edges in the network correspond to power lines between busses. The network is ordinarily sparse, with M a small multiple of N.

The power production at bus i is p_i, with negative values indicating consumption. For some busses, p_i is tightly controlled and deterministic in the relevant time horizon. Other busses have random p_i corresponding, for example, to variable consumption levels, that we treat as independent. Busses corresponding to wind farms have random power production levels with meaningfully large correlations. Our models contain one special bus S, called the slack bus, at which the power is $p_S = -\sum_{i\neq S} p_i$. The total power in the system is zero because transmission power losses are ignored in the DC approximation that we use.

The power at all busses can be represented by the vector $p = (p_F^{\mathsf{T}}, p_R^{\mathsf{T}}, p_S)^{\mathsf{T}}$ corresponding to fixed busses, ordinary random busses (including any correlated ones) and the slack bus. There are N_F fixed busses, $N_S = 1$ slack bus and $N_R = N - N_F - N_S$ random busses apart from the slack bus. We will use 1_R to denote a column vector of N_R ones, and I_R to denote the identity matrix of size N_R and similarly for 1_F and I_F.

The power p_i at bus i must satisfy the constraints

$$\underline{p}_i \leqslant p_i \leqslant \overline{p}_i. \tag{13}$$

The vector p has a Gaussian distribution, determined entirely by the random components $p_R \sim \mathcal{N}(\eta_R, \Sigma_{RR})$. Therefore in the present context, p_R is the Gaussian random variable $\boldsymbol{x}$ from Section 2. The fixed components satisfy $p_F = \eta_F$ and then the slack bus satisfies $p_S \sim \mathcal{N}(\eta_S, \Sigma_{SS})$ where $\eta_S = -1_R^{\mathsf{T}}\eta_R - 1_F^{\mathsf{T}}\eta_F$

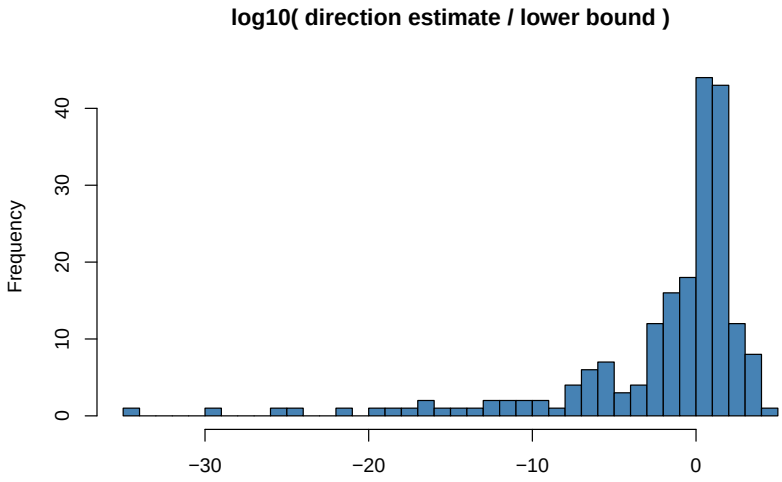


Figure 4: Results of 200 estimates of $\log_{10}(\mu_{\mathrm{AK}}/\underline{\mu})$ for varying high dimensional problems with nearly independent events.

and $\Sigma_{SS} = 1_R^{\mathsf{T}}\Sigma_{RR}1_R$. Because all of the randomness comes from p_R, we will abbreviate Σ_{RR} to Σ.

The node to node inductances in the network form a Laplacian matrix B where $B_{ij} \neq 0$ if busses i and j are connected with $B_{ii} = -\sum_{j\neq i} B_{ij}$ (up to rounding). The Laplacian is symmetric and has one eigenvalue of zero for a connected network. It has a pseudo-inverse B^+. We partition B and B^+ as follows

$$B = \begin{pmatrix} B_{RR} & B_{RF} & B_{RS} \\ B_{FR} & B_{FF} & B_{FS} \\ B_{SR} & B_{SF} & B_{SS} \end{pmatrix}, \quad \text{and} \quad B^+ = \begin{pmatrix} B^{RR} & B^{RF} & B^{RS} \\ B^{FR} & B^{FF} & B^{FS} \\ B^{SR} & B^{SF} & B^{SS} \end{pmatrix}.$$

We also group B^+ into three sets of columns via $B^+ = \begin{pmatrix} B^{\bullet R} & B^{\bullet F} & B^{\bullet S} \end{pmatrix}$.

The phase at bus i is denoted θ_i. In our DC approximation of AC power flow, the phases approximately satisfy $B\theta = p$. Given the power vector p, we take

$$\theta = B^+ p = \begin{pmatrix} B^{RR} & B^{RF} & B^{RS} \\ B^{FR} & B^{FF} & B^{FS} \\ B^{SR} & B^{SF} & B^{SS} \end{pmatrix} \begin{pmatrix} p_R \\ p_F \\ p_S \end{pmatrix}.$$

The phase constraints on the network are

$$|\theta_i - \theta_j| \leqslant \bar{\theta}_{ij}, \quad \text{for } i \neq j \text{ and } B_{ij} \neq 0. \tag{14}$$

In our examples, all $\bar{\theta}_{ij} = \bar{\theta}$ for a single value $\bar{\theta}$ such as $\pi/6$ or $\pi/4$.

Let $D \in \{-1, 1\}^{M\times N}$ be the incidence matrix. Each edge in the network is represented by one row of D with an entry of $+1$ for one of the busses on that

edge and -1 for the other. The phase constraints are $|D\theta| \leqslant \bar{\theta}$ componentwise. Now

$$D\theta = D\begin{pmatrix} B^{\bullet R} & B^{\bullet F} & B^{\bullet S}\end{pmatrix}\begin{pmatrix} p_R \\ p_F \\ p_S\end{pmatrix} = D\big(B^{\bullet R}p_R + B^{\bullet F}p_F + B^{\bullet S}p_S\big).$$

The constraint that $D\theta \leqslant \bar{\theta}$ for every edge ij can be written

$$DB^{\bullet R}p_R \leqslant \bar{\theta} - DB^{\bullet F}p_F - DB^{\bullet S}p_S.$$

Now $p_S = -1_R^{\mathsf{T}}p_R - 1_F^{\mathsf{T}}p_F$ and $p_F = \eta_F$, so the constraint on p_R is

$$D\big(B^{\bullet R} - B^{\bullet S}1_R^{\mathsf{T}}\big)p_R \leqslant \bar{\theta} - D\big(B^{\bullet F} - B^{\bullet S}1_F^{\mathsf{T}}\big)\eta_F. \tag{15}$$

We also have constraints $D\theta \geqslant -\bar{\theta}$ which can be written

$$D\big(B^{\bullet S}1_R^{\mathsf{T}} - B^{\bullet R})p_R\big) \leqslant \bar{\theta} + D\big(B^{\bullet F} - B^{\bullet S}1_F^{\mathsf{T}}\big)\eta_F. \tag{16}$$

Equations (15) and (16) supply $2M$ constraints on the random vector p_R. We have also the two slack bus constraints $p_S \leqslant \overline{p}_S$ and $-p_S \leqslant -\underline{p}_S$, that is

$$-1_R^{\mathsf{T}}p_R \leqslant \overline{p}_S + 1_F^{\mathsf{T}}\eta_F \quad \text{and} \quad 1_R^{\mathsf{T}}p_R \leqslant -\underline{p}_S + 1_F^{\mathsf{T}}\eta_F. \tag{17}$$

Finally, there are individual constraints on the random busses

$$p_R \leqslant \overline{p}_R, \quad \text{and} \quad -p_R \leqslant -\underline{p}_R,$$

componentwise.

When we combine all of the constraints on p_R, we get a matrix Γ with rows γ_j and a vector of upper bounds $\mathcal{K}$ with entries κ_j for which the constraints are $\Gamma \times p_R \leqslant \mathcal{K}$ componentwise. Here those matrices are

$$\Gamma = \begin{pmatrix} I_R \\ -I_R \\ 1_R^{\mathsf{T}} \\ -1_R^{\mathsf{T}} \\ D(B^{\bullet R} - B^{\bullet S}1_R^{\mathsf{T}}) \\ -D(B^{\bullet R} - B^{\bullet S}1_R^{\mathsf{T}})\end{pmatrix}, \quad \text{and} \quad \mathcal{K} = \begin{pmatrix} \overline{p}_R \\ -\underline{p}_R \\ -\underline{p}_S + 1_F^{\mathsf{T}}\eta_F \\ \overline{p}_S + 1_F^{\mathsf{T}}\eta_F \\ \bar{\theta} - D\big(B^{\bullet F} - B^{\bullet S}1_F^{\mathsf{T}}\big)\eta_F \\ \bar{\theta} + D\big(B^{\bullet F} - B^{\bullet S}1_F^{\mathsf{T}}\big)\eta_F\end{pmatrix}.$$

These are the linear constraints on p_R. There are $2M$ phase constraints and there are two constraints for all of the non-fixed busses, including the slack bus. These constraints can be turned into constraints Ω and $\mathcal{T}$ on a $\mathcal{N}(0, I_R)$ vector as described in Section 2.1.

$\bar{\omega}$	$\hat{\mu}$	se/$\hat{\mu}$	$\underline{\mu}$	$\bar{\mu}$
$\pi/4$	3.7×10^{-23}	0.0024	3.6×10^{-23}	4.2×10^{-23}
$\pi/5$	2.6×10^{-12}	0.0022	2.6×10^{-12}	2.9×10^{-12}
$\pi/6$	3.9×10^{-07}	0.0024	3.9×10^{-07}	4.4×10^{-07}
$\pi/7$	2.0×10^{-03}	0.0027	2.0×10^{-03}	2.4×10^{-03}

Table 2: Rare event estimates for the winter peak grid. $\bar{\omega}$ is the phase constraint, $\hat{\mu}$ is the ALOE estimate, se is the estimated standard error, $\underline{\mu}$ is the largest single event probability and $\bar{\mu}$ is the union bound.

5.2 Examples

We considered several model electrical grids included in the MATPOWER distribution (Zimmerman et al., 2011). In each case we modeled violations of the phase constraints, and used $n = 10{,}000$ samples. For some cases we found that, under our model, phase constraint violations were not rare events. In some other cases, the rare event probability was dominated by one single phase condition: $\underline{\mu} = \max_{j=1}^{J} P_j \approx \sum_{j=1}^{J} P_j = \bar{\mu}$. For cases like this there is no need for elaborate computation because we know μ is within a narrow interval $[\underline{\mu}, \bar{\mu}]$. The interesting cases were of rare events not dominated by a single failure mode. We investigate two of them.

The first is the Polish winter peak grid of 2383 busses. There were $d = 326$ random (uncontrolled) busses and $J = 5772$ phase constraints. We varied $\bar{\omega}$ as shown in Table 2. For $\bar{\omega} = \pi/7$ constraint violations are not very rare. At $\bar{\omega} = \pi/4$ they are quite rare. The estimated coefficient of variation is nearly constant over this range.

The second interesting case is the Pegase 2869 model of Fliscounakis et al. (2013). This has $d = 509$ uncontrolled busses and $J = 7936$ phase constraints. It is described as "power flow for a large part of the European system". The results are shown in Table 3. We include an unrealistically large bound $\bar{\omega} = \pi/2$ in that table, to test the limits of our approach. For $\bar{\omega} = \pi/2$, the standard error given is zero. One half-space was sampled 9408 times, another was sampled 592 times but in no instance were there two or more phase violations. The estimate reverts to the union bound. Getting 0 doubletons ($S = 2$) among $n = 10{,}000$ tries is compatible with the true probability of a doubleton being as high as $3/n$. Even if $T_1 = .9997$ and $T_2 = .0003$ then we would have $\mu = (1 - (3/2) \times 10^{-4})\bar{\mu}$ instead of $\bar{\mu}$. We return to this issue in the discussion.

In addition to the examples above we investigated IEEE case 14, IEEE case 300, and Pegase 1354, which were all dominated by one failure. We considered a system which included random and correlated wind power generators, but phase failure was not a rare event in that system. Pegase 13659 was too large for our computation. The Laplacian matrix has 37,250 rows and columns and we use the SVD to compute the generalized inverses we need. Pegase 9241 was large enough to be very slow and it did not have rare failures.

In our numerical tables we have used the plain sample variance of the impor-

$\bar{\omega}$	$\hat{\mu}$	se/$\hat{\mu}$	$\underline{\mu}$	$\bar{\mu}$
$\pi/2$	3.5×10^{-20}	0^*	3.3×10^{-20}	3.5×10^{-20}
$\pi/3$	8.9×10^{-10}	5.0×10^{-5}	7.7×10^{-10}	8.9×10^{-10}
$\pi/4$	4.3×10^{-06}	1.8×10^{-3}	3.5×10^{-06}	4.6×10^{-06}
$\pi/5$	2.9×10^{-03}	3.5×10^{-3}	1.8×10^{-03}	4.1×10^{-03}

Table 3: Rare event estimates for the Pegase 2869 model. The columns are as in Table 2. *The se was 0 because there were never two or more failures in any sample. See text for discussion.

tance sampled values to compute a standard error. Botev et al. (2015) note that the resulting standard error can be very inefficient and they propose a superior estimator. A naive implementation of their method would cost $O(J^4)$ for J linear constraints but they are able to reduce that cost to $O(J^3)$. We have not used that method here because with $J = 5772$ (Polish winter peak) or $J = 7936$ (Pegase 2869), even J^3 is too much to pay for a better variance estimate.

6 Discussion

We have introduced a version of mixture importance sampling for problems with multiple failure modes. The sample values are constrained to have at least one failure and we obtain bounded relative error.

The ALOE importance sampler is more accurate than a state of the art code for computing high dimensional Gaussian probabilities in our rare event setting, but not otherwise. It is also more reliable than the recent directional sampling method of Ahn and Kim (2018). That method gains accuracy by integrating over a randomly chosen line through the origin in $\mathbb{R}^d$, but it samples those lines uniformly and not by importance sampling. It is possible to combine the ideas, sampling $\boldsymbol{x}$ by ALOE and then integrating over the line defined by the unit vector $\boldsymbol{x}/\|\boldsymbol{x}\|$. Preliminary results show this to improve upon ALOE in the power grid problem however a full discussion would add too much length and detail to the present paper.

We have noticed two areas where ALOE can be improved. First, if we never see $S \geqslant 2$ concurrent rare events, ALOE will return the union bound $\hat{\mu} = \bar{\mu}$, with an estimated variance of zero. That variance estimate could be undesirable even when $\mu/\bar{\mu} \approx 1$. Because $S(\boldsymbol{x})$ is supported on $\{1, 1/2, \ldots, 1/J\}$ we can get an interval estimate of μ by putting a multinomial prior on this support set and using the posterior distribution given the sample. The solution in Botev et al. (2015) is attractive when J is not so large.

A second and related issue is that while $\bar{\mu} \geqslant \hat{\mu} \geqslant \bar{\mu}/J$ always occurs, it is possible to get $\hat{\mu} < \underline{\mu} = \max_{1 \leqslant j \leqslant J} P_j$. We have seen this in cases where $\bar{\mu} \approx \underline{\mu}$ because one of the P_j dominates all of the others combined. In such cases μ, $\bar{\mu}$ and $\underline{\mu}$ are all very close together and $\hat{\mu}$ has small relative standard deviation. Improving these two issues is outside the scope of this paper. They

are both things that happen in cases where we already have a very good idea of the magnitude of μ. The problem was much less severe for ALOE than it was for directional sampling. For instance, ALOE will not give $\hat{\mu} < \bar{\mu}/J$, while directional sampling can.

In large problems the algebra can potentially be reduced by ignoring the very rarest events and simply adding their probabilities to the estimate. This will provide a mildly conservative bound. There is also the possibility of exploiting many generalized upper and lower bounds on the probability of a union. See for instance the survey by Yang et al. (2014).

Acknowledgments

We thank Alan Genz for providing details of the `mvtnorm` package. We thank Yanbo Tang and Jeffrey Negrea for noticing that Lemma 1 could be proved by Cauchy-Schwarz, which is shorter than our original proof and also establishes necessity. Thanks also to Bert Zwart, Jose Blanchet, David Siegmund, Mark Huber, Zdravko Botev and an anonymous reviewer for pointers to the literature. ABO thanks the Center for Nonlinear Studies at Los Alamos for their hospitality while he visited.

References

Adler, R. J., J. Blanchet, and J. Liu (2008). Efficient simulation for tail probabilities of Gaussian random fields. In *Winter Simulation Conference, 2008*, pp. 328–336. IEEE.

Adler, R. J., J. H. Blanchet, and J. Liu (2012). Efficient Monte Carlo for high excursions of Gaussian random fields. *The Annals of Applied Probability 22*(3), 1167–1214.

Ahn, D. and K.-K. Kim (2018). Efficient simulation for expectations over the union of half-spaces. *ACM Transactions on Modeling and Computer Simulation (TOMACS) 28*(3), 23.

Asmussen, S. and P. W. Glynn (2007). *Stochastic simulation: algorithms and analysis*, Volume 57. Springer Science & Business Media.

Botev, Z. I. and P. L'Ecuyer (2015). Efficient probability estimation and simulation of the truncated multivariate student-t distribution. In *Winter Simulation Conference (WSC), 2015*, pp. 380–391. IEEE.

Botev, Z. I., M. Mandjes, and A. Ridder (2015). Tail distribution of the maximum of correlated gaussian random variables. In *Proceedings of the 2015 Winter Simulation Conference*, pp. 633–642. IEEE Press.

Chertkov, M., F. Pan, and M. G. Stepanov (2011). Predicting failures in power grids: The case of static overloads. *IEEE Transactions on Smart Grid 2*(1), 162–172.

Chertkov, M., M. Stepanov, F. Pan, and R. Baldick (2011). Exact and efficient algorithm to discover extreme stochastic events in wind generation over transmission power grids. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pp. 2174–2180.

Cornuet, J., J.-M. Marin, A. Mira, and C. P. Robert (2012). Adaptive multiple importance sampling. *Scandinavian Journal of Statistics 39*(4), 798–812.

Cranley, R. and T. N. L. Patterson (1976). Randomization of number theoretic methods for multiple integration. *SIAM Journal of Numerical Analysis 13*(6), 904–914.

Cunningham, S. W. (1969). Algorithm as 24: From normal integral to deviate. *Journal of the Royal Statistical Society. Series C 18*(3), 290–293.

Doucet, A. (2010). A note on efficient conditional simulation of Gaussian distributions. Technical report, University of British Columbia.

Elvira, V., L. Martino, D. Luengo, and M. F. Bugallo (2015a). Efficient multiple importance sampling estimators. *IEEE Signal Processing Letters 22*(10), 1757–1761.

Elvira, V., L. Martino, D. Luengo, and M. F. Bugallo (2015b). Generalized multiple importance sampling. *arXiv preprint arXiv:1511.03095*.

Fliscounakis, S., P. Panciatici, F. Capitanescu, and L. Wehenkel (2013). Contingency ranking with respect to overloads in very large power systems taking into account uncertainty, preventive, and corrective actions. *IEEE Transactions on Power Systems 28*(4), 4909–4917.

Frigessi, A. and C. Vercellis (1985). An analysis of Monte Carlo algorithms for counting problems. *Calcolo 22*(4), 413–428.

Genz, A. (2004). Numerical computation of rectangular bivariate and trivariate normal and t probabilities. *Statistics and Computing 14*(3), 251–260.

Genz, A. and F. Bretz (2009). *Computation of Multivariate Normal and t Probabilities*. Berlin: Springer-Verlag.

Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn (2017). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-6.

Hesterberg, T. C. (1988). *Advances in importance sampling*. Ph. D. thesis, Stanford University.

Kahn, H. and A. Marshall (1953). Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society of America 1*(5), 263–278.

Karp, R. M. and M. Luby (1983). Monte-carlo algorithms for enumeration and reliability problems. In *Foundations of Computer Science, 1983., 24th Annual Symposium on*, pp. 56–64. IEEE.

Kersulis, J., I. Hiskens, M. Chertkov, S. Backhaus, and D. Bienstock (2015, June). Temperature-based instanton analysis: Identifying vulnerability in transmission networks. In *2015 IEEE Eindhoven PowerTech*, pp. 1–6.

Lafortune, E. P. and Y. D. Willems (1993). Bidirectional path tracing. In *Proceedings of CompuGraphics*, pp. 95–104.

L'Ecuyer, P., M. Mandjes, and B. Tuffin (2009). Importance sampling and rare event simulation. In G. Rubino and B. Tuffin (Eds.), *Rare event simulation using Monte Carlo methods*, pp. 17–38. Chichester, UK: John Wiley & Sons.

Liu, J. S. (2001). *Monte Carlo strategies in scientific computing*. New York: Springer.

Miwa, T., A. Hayter, and S. Kuriki (2003). The evaluation of general non-centred orthant probabilities. *Journal of the Royal Statistical Society, Series B 65*(1), 223–234.

Naiman, D. Q. and C. E. Priebe (2001). Computing scan statistic p values using importance sampling, with applications to genetics and medical image analysis. *Journal of Computational and Graphical Statistics 10*(2), 296–328.

Niederreiter, H. (1972). On a number-theoretical integration method. *Aequationes Math 8*(3), 304–311.

Owen, A. B. (2013). *Monte Carlo theory, methods and examples.*

Owen, A. B. and Y. Zhou (2000). Safe and effective importance sampling. *Journal of the American Statistical Association 95*(449), 135–143.

Priebe, C. E., D. Q. Naiman, and L. M. Cope (2001). Importance sampling for spatial scan analysis: computing scan statistic p-values for marked point processes. *Computational statistics & data analysis 35*(4), 475–485.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sauer, P. W. and J. P. Christensen (1984). Active linear DC circuit models for power system analysis. *Electric machines and power systems 9*(2-3), 103–112.

Shi, J., D. O. Siegmund, and B. Yakir (2007). Importance sampling for estimating p values in linkage analysis. *Journal of the American Statistical Association 102*(479), 929–937.

Stott, B., J. Jardim, and O. Alsaç (2009). DC power flow revisited. *IEEE Transactions on Power Systems 24*(3), 1290–1300.

Van den Bergh, K., E. Delarue, and W. Dhaeseleer (2014). DC power flow in unit commitment models. Technical Report WP EN2014-12, KU Leuven.

Veach, E. and L. Guibas (1994, June 13–15). Bidirectional estimators for light transport. In *5th Annual Eurographics Workshop on Rendering*, pp. 147–162.

Yang, J., F. Alajaji, and G. Takahara (2014). A short survey on bounding the union probability using partial information. Technical report, University of Toronto.

Zimmerman, R. D., C. E. Murillo-Sánchez, and R. J. Thomas (2011). MATPOWER: steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on power systems 26*(1), 12–19.

Appendix: Proofs

Analysis of self-normalized importance sampling

Here we show that self-normalized importance sampling cannot attain the smallest variances in a rare event setting. We build on a remark by Hesterberg (1988, Chapter 2) and follow a derivation from Owen (2013, Chapter 9). The key problem is that the optimal self-normalized importance sampler for a rare event places only $1/2$ of its probability in the rare event.

Hesterberg (1988, Chapter 2) quoting Kahn and Marshall (1953) notes that the optimal self-normalized importance sampling density q for estimating $\mu = \mathbb{E}(f(\boldsymbol{x}))$ when $\boldsymbol{x} \sim p$ is proportional to $|f(\boldsymbol{x})-\mu|p(\boldsymbol{x})$. Taking $f(\boldsymbol{x}) = \mathbf{1}\{\boldsymbol{x} \in A\}$, we get $q(\boldsymbol{x}) = |\mathbf{1}\{\boldsymbol{x} \in A\}-\mu|p(\boldsymbol{x})/c$ for some $c > 0$. Solving $q(A \cup A^c) = 1$ yields $c = 2\mu(1-\mu)$ and then $q(A) = q(A^c) = 1/2$. The self-normalized importance sampler is a ratio estimate $\sum_i f(\boldsymbol{x}_i)p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i) \,/\, \sum_i p(\boldsymbol{x}_i)/q(\boldsymbol{x}_i)$. Its variance is asymptotic to σ_q^2/n for

$$\sigma_q^2 = \mathbb{E}_q\Big(\frac{p(\boldsymbol{x})^2(f(\boldsymbol{x})-\mu)^2}{q(\boldsymbol{x})^2}\Big)$$

Half of the time q places $\boldsymbol{x} \in A$ yielding $p(\boldsymbol{x})/q(\boldsymbol{x}) = (1-\mu)^{-1}/c$ and half of the time $\boldsymbol{x} \in A^c$ with $p(\boldsymbol{x})/q(\boldsymbol{x}) = \mu^{-1}/c$. Thus

$$\sigma_q^2 = \frac{1}{2}\Big[\frac{1}{c^2} + \frac{1}{c^2}\Big] = 4\mu^2(1-\mu)^2.$$

The best possible asymptotic coefficient of variation for SNIS is then $2(1-\mu)/\sqrt{n} \approx 2/\sqrt{n}$ for rare events.

Proof of Theorem 2

Our motivating problem involves probabilities defined by Gaussian content of half-spaces. The approach generalizes to estimating the probability of the union of any finite set of events. We can also consider a countable number of events when the union bound is finite; see remarks below. We will assume that each event has positive probability, but that condition can also be weakened to a positive union bound, as described at the end of this section.

For definiteness, we define our sets in terms of indicator functions of a random variable $\boldsymbol{x} \in \mathbb{R}^d$ with probability density p. The same formulas work for general sample spaces and the density can be with respect to an arbitrary base measure.

We cast our notation into this more general setting as follows. For $J \geqslant 1$, and $j = 1, \ldots, J$, let the subset $H_j \subset \mathbb{R}^d$ define both the event $H_j = \mathbf{1}\{\boldsymbol{x} \in H_j\}$ and the indicator function $H_j(\boldsymbol{x}) = 1\{\boldsymbol{x} \in H_j\}$. For $u \subseteq \{1, \ldots, J\}$ we let $H_u = \cup_{j \in u} H_j$ and $H_u(\boldsymbol{x}) = \max_{j \in u} H_j(\boldsymbol{x})$, with $H_\varnothing(\boldsymbol{x}) = 0$. As before $P_u = \mathbb{E}(H_u(\boldsymbol{x}))$, the number of events is $S(\boldsymbol{x}) = \sum_{j=1}^J H_j(\boldsymbol{x})$ and $\Pr(S = s) = T_s$.

Recall that we use $-u$ for complements with respect to $1{:}J$, especially within subscripts, and $H_u^c(\boldsymbol{x})$ for the complementary outcome $1 - H_u(\boldsymbol{x})$. Then $H_u(\boldsymbol{x})H_{-u}^c(\boldsymbol{x})$ describes the event where $\boldsymbol{x}_j \in H_j$ if and only if $j \in u$.

If $P_j > 0$, then the distribution q_j of $\boldsymbol{x}$ given H_j is well defined: $q_j(\boldsymbol{x}) = p(\boldsymbol{x})H_j(\boldsymbol{x})/P_j$. If $\min_j P_j > 0$ then we can define the mixture distribution

$$q_{\alpha^*} = \sum_{j=1}^J \alpha_j^* q_j, \quad \alpha_j^* = P_j/\bar{\mu}, \quad \bar{\mu} = \sum_{j=1}^J P_j. \tag{18}$$

For $n \geqslant 1$, our estimator of $\mu = \Pr(S(\boldsymbol{x}) > 0)$ is

$$\hat{\mu}_{\alpha^*} = \frac{\bar{\mu}}{n} \sum_{i=1}^n \frac{1}{S(\boldsymbol{x}_i)}, \quad \boldsymbol{x}_i \overset{\text{iid}}{\sim} q_{\alpha^*}. \tag{19}$$

In this section, some equations include both randomness due to $\boldsymbol{x}_i \sim q_{\alpha^*}$ and randomness due to $\boldsymbol{x} \sim p$. For section only, we use $\Pr_*$, $\mathbb{E}_*$ and Var_* when the randomness is from observations $\boldsymbol{x}_i \sim q_{\alpha^*}$, while $\mathbb{E}$, Pr and Var are with respect to $\boldsymbol{x} \sim p$.

Theorem 2. *If* $1 \leqslant J < \infty$ *and* $\min_j P_j > 0$ *and* $n \geqslant 1$, *then* $\hat{\mu}_{\alpha^*}$ *defined by* (18) *satisfies* $\Pr_*(\bar{\mu}/J \leqslant \hat{\mu}_{\alpha^*} \leqslant \bar{\mu}) = 1$,

$$\mathbb{E}_*(\hat{\mu}_{\alpha^*}) = \mu, \tag{20}$$

and

$$\mathrm{Var}_*(\hat{\mu}_{\alpha^*}) = \frac{1}{n}\left(\bar{\mu} \sum_{s=1}^J \frac{T_s}{s} - \mu^2\right) \leqslant \frac{\mu(\bar{\mu} - \mu)}{n}. \tag{21}$$

Proof of Theorem 2. Let $H(\boldsymbol{x}) = \max_{1\leqslant j\leqslant J} H_j(\boldsymbol{x})$ and $H = \{\boldsymbol{x} \mid H(\boldsymbol{x}) = 1\}$. If $\boldsymbol{x}_i \sim q_{\alpha^*}$, then $\boldsymbol{x}_i \in H$ always holds. Then $1 \leqslant S(\boldsymbol{x}_i) \leqslant J$ holds establishing the bounds on $\hat{\mu}_{\alpha^*}$. Next

$$\mathbb{E}_*\left(\Big(\sum_{j=1}^J H_j(\boldsymbol{x}_1)\Big)^{-1}\right) = \sum_{\ell=1}^J \frac{P_\ell}{\bar{\mu}} \int_H \frac{H_\ell(\boldsymbol{x}) P_\ell^{-1} p(\boldsymbol{x})}{\sum_{j=1}^J H_j(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} = \frac{1}{\bar{\mu}}\int_H p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \frac{\mu}{\bar{\mu}},$$

establishing (20).

Because $\hat{\mu}_{\alpha^*}$ is unbiased its variance is

$$\frac{1}{n}\left(\bar{\mu}^2 \mathbb{E}_*\left(\left(\frac{H(\boldsymbol{x}_1)}{\sum_{j=1}^J H_j(\boldsymbol{x}_1)}\right)^2\right) - \mu^2\right). \tag{22}$$

Next

$$\begin{aligned}
\mathbb{E}_*\Big(\Big(\frac{H(\boldsymbol{x}_1)}{\sum_j H_j(\boldsymbol{x}_1)}\Big)^2\Big) &= \sum_{j=1}^J \frac{P_j}{\bar{\mu}} \int_H \Big(\frac{H(\boldsymbol{x})}{1+\sum_{\ell\neq j} H_\ell(\boldsymbol{x})}\Big)^2 \frac{p(\boldsymbol{x}) H_j(\boldsymbol{x})}{P_j}\,\mathrm{d}\boldsymbol{x} \\
&= \bar{\mu}^{-1} \int_H \frac{p(\boldsymbol{x})}{\sum_{j=1}^J H_j(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} \\
&= \bar{\mu}^{-1} \sum_{|u|>0} \frac{1}{|u|} \int H_u(\boldsymbol{x}) H^c_{-u}(\boldsymbol{x}) p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\
&= \bar{\mu}^{-1} \sum_{s=1}^J \frac{1}{s} T_s,
\end{aligned}$$

which, with (22), establishes the equality in (21). Next

$$\bar{\mu}^{-1} \sum_{s=1}^J \frac{T_s}{s} \leqslant \bar{\mu}^{-1} \sum_{s=1}^J T_s = \bar{\mu}^{-1}(1-T_0) = \bar{\mu}^{-1}\mu.$$

Finally $\bar{\mu}^2(\bar{\mu}^{-1}\mu) - \mu^2 = (\bar{\mu}-\mu)\mu$, establishing the upper bound in (21). □

We can generalize the previous theorem to higher moments. Our estimate is an average of $\bar{\mu}/S(\boldsymbol{x}_i)$. The k'th moment of this quantity is

$$\begin{aligned}
\mathbb{E}_*\Big(\Big(\frac{\bar{\mu}}{S(\boldsymbol{x}_1)}\Big)^k\Big) &= \bar{\mu}^k \sum_{j=1}^J \alpha_j \int_H S(\boldsymbol{x})^{-k} q_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\
&= \bar{\mu}^k \sum_{j=1}^J \frac{P_j}{\bar{\mu}} \int_H S(\boldsymbol{x})^{-k} \frac{p(\boldsymbol{x}) H_j(\boldsymbol{x})}{P_j}\,\mathrm{d}\boldsymbol{x} \\
&= \bar{\mu}^{k-1} \int_H S(\boldsymbol{x})^{1-k}\,\mathrm{d}\boldsymbol{x} \\
&= \sum_{s=1}^J T_s \Big(\frac{\bar{\mu}}{s}\Big)^{k-1}.
\end{aligned}$$

Remark 1. Suppose that one of the $P_j = 0$ but $\bar{\mu} > 0$. In this case q_j is not well defined. However q_{α^*} places probability 0 on q_j, so we may delete the q_j component without changing the algorithm and then sampling from q_{α^*} is well defined.

Remark 2. Next suppose that there are infinitely many events, one for each $j \in \mathbb{N}$. If $\bar{\mu} \in (0, \infty)$, then q_{α^*} is well defined. The same proof goes through, only now sums over $1{:}J$ must be replaced by sums over $\mathbb{N}$.

Proof of Lemma 1

From the Cauchy-Schwarz inequality,

$$1 - \mathbb{E}(S)\mathbb{E}(S^{-1}) = \mathrm{Cov}(S, S^{-1}) \geqslant -\sqrt{\mathrm{Var}(S)\mathrm{Var}(S^{-1})}. \tag{23}$$

Now $\mathrm{Var}(S) \leqslant (J-1)^2/4$ and $\mathrm{Var}(S^{-1}) \leqslant (1-J^{-1})^2/4$ because the support of S is in $[1, J]$. Therefore

$$\mathbb{E}(S)\mathbb{E}(S^{-1}) \leqslant 1 + \frac{(J-1)(1-J^{-1})}{4} = \frac{J + J^{-1} + 2}{4}.$$

Finally, the unique distribution for which $\mathrm{Var}(S)$, $\mathrm{Var}(S^{-1})$ and $-\mathrm{Corr}(S, S^{-1})$ all attain their maxima is $\mathbb{U}\{1, J\}$. $\square$

Generalization

The lemma generalizes. If $\Pr(a \leqslant X \leqslant b) = 1$ for $0 < a \leqslant b < \infty$ then the same argument yields $\mathbb{E}(X)\mathbb{E}(X^{-1}) \leqslant (a/b + b/a + 2)/4$.