Estimation and Asymptotics for Buffered Probability of Exceedance

Alexander Mafusalov
University of Florida
mafusalov@ufl.edu

Alexander Shapiro
Georgia Institute of Technology
ashapiro@isye.gatech.edu

Stan Uryasev
University of Florida
uryasev@ufl.edu

This draft: October 2015

RESEARCH REPORT 2015-5
Risk Management and Financial Engineering Lab
Department of Industrial and Systems Engineering
303 Weil Hall, University of Florida, Gainesville, FL 32611.
E-mail: *mafusalov@ufl.edu, ashapiro@isye.gatech.edu, uryasev@ufl.edu*

Abstract

The paper studies statistical properties of empirical (sample) estimates of the Buffered Probability of Exceedance (bPOE). The estimation procedure is based on one dimensional minimization representation of the bPOE. We investigate convergence rates and asymptotic properties of the suggested estimation procedures. Theoretical predictions are validated with numerical experiments.

Key words Buffered Probability of Exceedance, Conditional Value-at-Risk, Sample Average Approximation, large samples statistical inference, rare events, minimum volume ellipsoid.

1 Introduction

Buffered Probability of Exceedance (bPOE) for a random value is a counterpart to the Probability of Exceedance (POE). The bPOE notion was introduced and studied in Mafusalov and Uryasev [7] and in Norton and Uryasev [8]. For a specified threshold, bPOE equals the probability of an upper tail of the distribution, such that the average of this tail coincides with the threshold. The bPOE of a random value is similar to POE and it is an upper bound for POE because it includes all outcomes exceeding the threshold, as well as some outcomes below the threshold. The outcomes below the threshold form the so called buffer, therefore, bPOE is a buffered POE.

The bPOE concept is an extension of the so called Buffered Probability of Failure (bPOF) suggested by Rockafellar [9] and explored by Rockafellar and Royset [11]. The bPOF equals one minus inverse at point zero of Conditional Value-at-Risk (CVaR) of a random value. See definition and properties of CVaR at Rockafellar and Uryasev [12]. Similar, bPOE for some random value, X, at a threshold $x \in \mathbb{R}$ equals $1-\alpha$, where

$$\mathsf{CVaR}_{\alpha}(X) = x.$$

The standard POE is very popular in various engineering applications. For instance, nuclear engineering considers probability that radiation release will exceed specified level and structural reliability analysis considers probability that load exceeds some threshold. Although POE is very popular and it is included in government regulations, it has some major shortcomings. From conceptual point of view, the threshold in POE provides a low bound on tail outcomes exceeding this threshold. The POE does not provide information of the magnitude of outcomes exceeding the threshold. Also, POE has some troublesome mathematical properties for discretely distributed random values, which are typically obtained from sample data. POE is discontinuous with respect to the threshold, which prevents using standard sensitivity analysis based on derivatives. Also, POE is difficult to optimize because optimization problems for POE are usually reduced to difficult Mixed-Integer optimization involving many binary variables.

The bPOE is a nice mathematical function. It is continuous in threshold z (may be except one point) and quasi-convex in X. Moreover, bPOE minimization problem can be reduced to convex and even linear programming. Mafusalov and Uryasev [7] provide detail description of mathematical properties of bPOE and various optimization problem statements.

This paper studies statistical properties of empirical (sample) estimates of bPOE. The estimators are based on one-dimension minimization representation of bPOE suggested in Mafusalov and Uryasev [7] and in Norton and Uryasev [8]. We also study asymptotic convergence of the suggested estimation algorithms. The theoretical results are validated with numerical experiments.

2 Statistical Properties of Buffered Probability Estimates

For $\alpha \in (0,1)$ Conditional Value-at-Risk of a random variable X is defined as[1]

$$\mathsf{CVaR}_{\alpha}(X) := \inf_{t\in\mathbb{R}} \left\{ t + (1-\alpha)^{-1}\mathbb{E}[X-t]_{+} \right\}. \tag{2.1}$$

We assume that $\mathbb{E}|X| < \infty$, and hence the expectation in (2.1) is well defined and finite valued. For $\alpha = 0$, $\mathsf{CVaR}_0(X) = \mathbb{E}[X]$ and $\mathsf{CVaR}_{\alpha}(X)$ tends to the essential supremum[2] of X as $\alpha \uparrow 1$, so

[1] We use notation $[a]_+ := \max\{0, a\}$ for $a \in \mathbb{R}$.

[2] The essential supremum $\operatorname{ess\,sup}(X)$ can be $+\infty$ if the random variable X is unbounded.

we define $\mathsf{CVaR}_1(X) := \operatorname{ess\,sup}(X)$. Let $F_X(x) = \mathsf{Prob}(X \leq x)$ be the cumulative distribution function of X and

$$q^-_\alpha(X) := \inf\{t : F_X(x) \geq \alpha\},\ q^+_\alpha(X) := \sup\{t : F_X(x) \leq \alpha\},$$

be the left side and right side quantiles of X. If $q^-_\alpha(X) = q^+_\alpha(X)$ we simply denote it by $q_\alpha(X)$. It is well known that for $\alpha \in (0,1)$ the minimum in the right hand side of (2.1) is attained for any $t \in [q^-_\alpha(X), q^+_\alpha(X)]$.

Denote $\bar{q}_\alpha(X) := \mathsf{CVaR}_\alpha(X)$. Consider the equation

$$x = \mathsf{CVaR}_\alpha(X). \tag{2.2}$$

It follows from the representation

$$\mathsf{CVaR}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 q^-_\tau(X)d\tau, \tag{2.3}$$

that $\mathsf{CVaR}_\alpha(X)$ is continuous and monotonically increasing in $\alpha \in [0,1)$, i.e., $\mathsf{CVaR}_{\alpha_1}(X) < \mathsf{CVaR}_{\alpha_2}(X)$ for any $0 \leq \alpha_1 < \alpha_2 < 1$. Hence equation (2.2) has unique solution $\alpha = \bar{q}_x^{-1}(X)$ for $\mathbb{E}[X] \leq x < \operatorname{ess\,sup}(X)$. The buffered probability of exceedance of a random variable X is defined as

$$\bar{p}_x(X) := \begin{cases} 1 - \bar{q}_x^{-1}(X) & \text{if} \quad \mathbb{E}[X] < x < \operatorname{ess\,sup}(X), \\ 1 & \text{if} \quad x \leq \mathbb{E}[X], \\ 0 & \text{if} \quad x \geq \operatorname{ess\,sup}(X). \end{cases} \tag{2.4}$$

That is,

$$\mathsf{CVaR}_{1-\bar{p}_x(X)}(X) = x, \text{ when } \mathbb{E}[X] \leq x < \operatorname{ess\,sup}(X).$$

Consider the following representation of the buffered probability of exceedance of a random variable X (cf., [7, Proposition 1]):

$$\bar{p}_x(X) = \begin{cases} \inf_{a \geq 0} \mathbb{E}[a(X-x)+1]_+ & \text{if} \quad x < \operatorname{ess\,sup}(X), \\ 0 & \text{if} \quad x \geq \operatorname{ess\,sup}(X). \end{cases} \tag{2.5}$$

Consider $\Psi(a, X) := [a(X-x)+1]_+$ and $\psi(a) := \mathbb{E}[\Psi(a, X)]$. Note that $\Psi(a, X)$ and hence $\psi(a)$ are convex functions of a. For $\mathbb{E}[X] < x < \operatorname{ess\,sup}(X)$ the set of minimizers $\arg\min_{a \geq 0} \psi(a)$ forms a closed interval $[a_1, a_2]$, where

$$a_1 = 1/(x - q^-_\alpha(X)) \ \text{ and } \ a_2 = 1/(x - q^+_\alpha(X)), \tag{2.6}$$

with α is defined by equation (2.2). In particular if the quantile $q_\alpha(X)$ is unique, i.e., $q_\alpha(X) = q^-_\alpha(X) = q^+_\alpha(X)$, then

$$\bar{a} = 1/(x - q_\alpha(X)) \tag{2.7}$$

is the unique minimizer of the right hand side of (2.5). For $\alpha \in (0,1)$ we have that $\mathsf{CVaR}_\alpha(X) > q^-_\alpha(X)$, and hence the numbers a_1 and a_2 are positive when $\mathbb{E}[X] < x$. When $x < \mathbb{E}[X]$, the minimizer in (2.5) is $\bar{a} = 0$, and $\bar{p}_X(x) = 1$. When $x < \operatorname{ess\,sup}(X)$ we have that $X < x$ w.p.1, and hence $\inf_{a \geq 0} \psi(a) = 0 = \bar{p}_X(x)$. When $x = \operatorname{ess\,sup}(X)$,

$$\inf_{a \geq 0} \psi(a) = \mathsf{Prob}(X = x), \tag{2.8}$$

and hence $\inf_{a \geq 0} \psi(a) > \bar{p}_x(X)$ if $\mathsf{Prob}(X = x) > 0$.

Let $X^1, ..., X^N$ be an iid random sample of X. By $\bar{X} := N^{-1}(X^1 + ... + X^N)$ we denote average of the sample. Consider

$$\hat{\psi}_N(a) := \frac{1}{N}\sum_{j=1}^{N} \Psi(a, X^j) = \frac{1}{N}\sum_{j=1}^{N}[a(X^j - x) + 1]_+.$$

That is, $\hat{\psi}_N(a)$ is the empirical (sample average) estimate of the expectation $\psi(a)$. Then a natural estimator of $\bar{p}_x(X)$ is obtained by replacing the probability distribution of X with its empirical estimate. Hence we consider the following estimator of $\bar{p}_x(X)$:

$$\hat{p}_N(x) = \begin{cases} \inf_{a\geq 0} \hat{\psi}_N(a) & \text{if} \quad x < \max\{X^1, ..., X^N\}, \\ 0 & \text{if} \quad x \geq \max\{X^1, ..., X^N\}. \end{cases} \tag{2.9}$$

By $\hat{a}_N$ we denote a minimizer of the right hand side of (2.9). This minimizer can be computed as above using the empirical distribution given by the considered sample. That is, $\hat{a}_N$ can be any number in the interval $[\hat{a}_{1N}, \hat{a}_{2N}]$, where $\hat{a}_{1N}$ and $\hat{a}_{2N}$ are obtained by replacing in equation (2.6) the respective quantiles by their sample estimates. Note that $\hat{p}_N(x) = 1 - \hat{\alpha}$, where $\hat{\alpha}$ is computed by using equation (2.2) with $\mathsf{CVaR}_\alpha(X)$ replaced by its empirical estimate. That is

$$x = \widehat{\mathsf{CVaR}}_{\hat{\alpha}N} \tag{2.10}$$

where

$$\widehat{\mathsf{CVaR}}_{\alpha N} = \inf_{t\in\mathbb{R}} \left\{ t + (1-\alpha)^{-1}N^{-1}\sum_{i=1}^{N}[X_i - t]_+ \right\}. \tag{2.11}$$

The minimizer in the right hand side of (2.11) is given by the empirical quantile $\bar{t} = \hat{q}_\alpha$. Note that if $x \leq \bar{X}$, then $\hat{a}_N = 0$ and $\hat{p}_N(x) = 1$.

We can view the right hand side of (2.9) as the Sample Average Approximation (SAA) of the stochastic problem in the right hand side of (2.5). Hence some standard results of the statistical inference of the SAA problems can be applied. In particular we have the following results.

Theorem 2.1 *Suppose that $\mathbb{E}|X| < \infty$ and let $\alpha = \bar{q}_x^{-1}(X)$ for $\mathbb{E}[X] \leq x < \operatorname{ess\,sup}(X)$. Then the following holds.* (i) *The estimator $\hat{p}_N(x)$ converges to $\bar{p}_x(X)$ w.p.1 uniformly on any interval $[c, d]$ such that $\mathbb{E}[X] < c \leq d < \operatorname{ess\,sup}(X)$, i.e.,*

$$\lim_{N\to\infty} \sup_{x\in[c,d]} \left|\hat{p}_N(x) - \bar{p}_x(X)\right| = 0, \ w.p.1. \tag{2.12}$$

(ii) *If $\mathbb{E}[X] \leq x < \operatorname{ess\,sup}(X)$, then the bias $\mathbb{E}\left[\hat{p}_N(x)\right] - \bar{p}_X(x)$ of the estimator $\hat{p}_N(x)$ is* negative, *and this bias is monotonically decreasing, i.e.,*

$$\mathbb{E}\left[\hat{p}_N(x)\right] \leq \mathbb{E}\left[\hat{p}_{N+1}(x)\right] \leq \bar{p}_x(X). \tag{2.13}$$

(iii) *If $\mathbb{E}[X] < x < \operatorname{ess\,sup}(X)$, the quantile $q_\alpha(X)$ is unique and variance*

$$\sigma^2(x) := \operatorname{Var}\left\{[\bar{a}(X - x) + 1]_+\right\} \tag{2.14}$$

is finite ($\bar{a}$ is defined in (2.7)*), then*

$$\hat{p}_N(x) = \frac{1}{N}\sum_{j=1}^{N}[\bar{a}(X^j - x) + 1]_+ + o_p(N^{-1/2}), \tag{2.15}$$

and $N^{1/2}\left(\hat{p}_N(x) - \bar{p}_x(X)\right)$ converges in distribution to normal $\mathcal{N}(0, \sigma^2(x))$.

Proof Let us observe that the empirical estimate $\widehat{\mathsf{CVaR}}_{\alpha N}$ converges to $\mathsf{CVaR}_\alpha(X)$ w.p.1 uniformly in $\alpha \in [\gamma_1, \gamma_2]$, where $0 < \gamma_1 \leq \gamma_2 < 1$. Indeed, by using a uniform Law of Large Numbers (e.g., [13, Theorem 7.53]) we have that $N^{-1}\sum_{j=1}^N [X^j - t]_+$ converges to $\mathbb{E}[X-t]_+$ w.p.1 uniformly in t on any finite interval. For $\alpha \in [\gamma_1, \gamma_2]$ we can restrict the minimization in t in (2.1) and (2.11) to a finite interval, and hence the uniform convergence of $\widehat{\mathsf{CVaR}}_{\alpha N}$ to $\mathsf{CVaR}_\alpha(X)$ follows. Since $\mathsf{CVaR}_\alpha(X)$ is continuous and monotonically increasing, this implies uniform convergence w.p.1 in $x \in [c, d]$ of the empirical estimate of the inverse function $\bar{q}_x^{-1}(X)$, and hence (2.12) follows.

For proof of assertions (ii) and (iii) we can refer to [13, Proposition 5.6] and [13, Theorem 5.7], respectively, by using representation (2.5). ∎

Remark 1 Of course it follows from the assertion (i) of the above theorem that $\hat{p}_N(x)$ converges to $\bar{p}_x(X)$ w.p.1 for any $\mathbb{E}[X] < x < \operatorname{ess\,sup}(X)$. For $x \leq \mathbb{E}[X]$ the set of minimizers in (2.5) is bounded (unless X is constant), and includes $\bar{a} = 0$. Note that the corresponding function $\Psi(a, X)$ is convex in a. We can apply [13, Theorem 5.4] to conclude that $\hat{p}_N(x)$ converges to $\bar{p}_x(X)$ w.p.1. If $x \geq \operatorname{ess\,sup}(X)$ and $\mathsf{Prob}(X = x) = 0$, then both $\bar{p}_x(X)$ and $\hat{p}_N(x)$ are zeros. Finally consider the case of $x = \operatorname{ess\,sup}(X)$ and $\mathsf{Prob}(X = x) > 0$. Then probability that at least one of X_i is equal to x, and hence $x = \max\{X_1, ..., X_N\}$, tends to one. In that case $\hat{p}_N(x)$ converges to $\bar{p}_x(X)$ in probability rather than w.p.1.

Remark 2 Variance $\sigma^2(x)$ can be estimated by

$$\hat{\sigma}^2(x) := \frac{1}{N-1}\sum_{j=1}^N \Big([\hat{a}_N(X^j - x) + 1]_+ - \hat{p}_N(x)\Big)^2. \tag{2.16}$$

When the (true) $\alpha > 0$ we have that $x > \mathbb{E}[X]$. However if α is close to 0, and hence x is close to $\mathbb{E}[X]$, it can happen that $x \leq \bar{X}$ in which case $\hat{p}_N(x) = 1$. Therefore for α close to 0, a better approximation of the distribution of $\hat{p}_N(x)$ will be the mixture[3] of distributions $\delta(1)$ and $\mathcal{N}(0, \sigma^2(x))$ with the respective weights ρ and $1 - \rho$, where $\rho := \mathsf{Prob}(x \leq \bar{X})$. By the CLT the distribution of $\bar{X}$ can be approximated (under standard regularity conditions) by the normal distribution $\mathcal{N}(\mu, \nu^2/N)$, where $\mu := \mathbb{E}[X]$ and $\nu^2 := \mathrm{Var}(X)$. Consequently the probability ρ can be approximated by[4] $1 - \Phi\big(\sqrt{N}(x-\mu)/\nu\big)$. In particular, if $\sqrt{N}(x-\mu)/\nu$ is greater than, say, three, then the probability ρ can be negligibly small and the normal distribution $\mathcal{N}(0, \sigma^2(x))$ could give a good approximation of the distribution of $\hat{p}_N(x)$.

When α is close to one, and hence x is close to[5] $\operatorname{ess\,sup}(X)$ and $F_X(x)$ is close to one, it can happen that $x \geq \max\{X^1, ..., X^N\}$ in which case $\hat{p}_N(x) = 0$. The probability

$$\varrho := \mathsf{Prob}\left(x \geq \max\{X^1, ..., X^N\}\right) = \prod_{j=1}^N \mathsf{Prob}(X^j \leq x) = (1-p)^N \approx e^{-Np}, \tag{2.17}$$

where $p := 1 - F_X(x) = \mathsf{Prob}(X > x)$. Therefore we will need a sample size of order $p^{-1}\ln\varepsilon^{-1}$ to make this probability ϱ less than $\varepsilon > 0$. If the probability ϱ is not small, then we can use the mixture of distributions $\delta(0)$ and $\mathcal{N}(0, \sigma^2(x))$, with the respective weights ϱ and $1 - \varrho$, as an approximation of the distribution of $\hat{p}_N(x)$.

[3] $\delta(y)$ denotes probability measure of mass 1 at the point y.

[4] $\Phi(\cdot)$ denotes the cumulative distribution function of standard normal distribution.

[5] If $\operatorname{ess\,sup}(X) = +\infty$, this means that x is large.

The right hand side of (2.15) gives a first order expansion of the estimator $\hat{p}_N(x)$. Under stronger assumptions it is possible to derive a second order term in the corresponding expansion (cf., [13, p.338]). That is, under appropriate regularity conditions,

$$\hat{p}_N(x) - \hat{\psi}_N(\bar{a}) = N^{-1} \inf_{\tau \in \mathbb{R}} \left\{ \tau V + \tfrac{1}{2}\tau^2 \psi''(\bar{a}) \right\} + o_p(N^{-1}) \tag{2.18}$$

$$= -\frac{V^2}{2N\psi''(\bar{a})} + o_p(N^{-1}). \tag{2.19}$$

where $V \sim N(0, \gamma^2)$ with

$$\gamma^2 = \mathrm{Var}\left(\frac{\partial \Psi(\bar{a}, X)}{\partial a} \right) = \mathbb{E}\left[\frac{\partial \Psi(\bar{a}, X)}{\partial a} \right]^2. \tag{2.20}$$

Note that $\psi'(\bar{a}) = \mathbb{E}\left[\frac{\partial \Psi(\bar{a}, X)}{\partial a} \right] = 0$ by optimality of $\bar{a}$.

We have that

$$\mathbb{E}[\hat{\psi}_N(\bar{a})] = \psi(\bar{a}) = \bar{p}_x(X),$$

and hence the bias can be approximated as

$$\mathbb{E}[\hat{p}_N(x)] - \bar{p}_x(X) = -\frac{\gamma^2}{2N\psi''(\bar{a})} + o(N^{-1}). \tag{2.21}$$

Assuming that random variable X has continuous probability density function $f(\cdot)$, we have

$$\psi'(a) = \mathbb{E}\left[\frac{\partial \Psi(a, X)}{\partial a} \right] = \int_{-1/a}^{+\infty} t f(t + x) dt,$$

and hence

$$\psi''(\bar{a}) = \frac{f(x - 1/\bar{a})}{\bar{a}^3} = (x - q_\alpha(X))^3 f(q_\alpha(X)).$$

3 Rare events

Consider now the case where probability $p := P(X \geq x)$ is very small, say of order 10^{-5} or smaller. In that case, given a sample $X^1, ..., X^N$, it will be difficult to employ the corresponding estimator $\hat{p}_N(x)$ in a straightforward way. Recall that $\hat{p}_N(x)$ is zero if x is greater than $\max\{X^1, ..., X^N\}$ (see (2.9)). The probability that at least one of the samples $X^1, ..., X^N$ is greater than x is $1 - (1 - p)^N \approx 1 - e^{-Np}$ (see (2.17)). That is, in order to have a reasonable probability for $\max\{X^1, ..., X^N\}$ to be less than x, i.e., for the estimator $\hat{p}_N(x)$ to be greater than zero, we will need a sample size of order of millions. This, of course, could be practically infeasible.

Suppose that X has normal distribution $\mathcal{N}(\mu, \nu^2)$ and $x = \mu + k\nu$, where $k \geq 4$, say. Then the probability $P(X \geq x) = 1 - \Phi(x)$ is very small. For example for $x = \mu + 4\nu$ the probability $p = 1 - \Phi(4) = 3 \times 10^{-5}$. Nevertheless we can proceed as follows. It is known that in the case of normal distribution,

$$\mathsf{CVaR}_\alpha(X) = \mu + \frac{\nu}{(1 - \alpha)\sqrt{2\pi}} e^{-z_\alpha^2/2}, \tag{3.1}$$

where $z_\alpha = \Phi^{-1}(\alpha)$. In that case first we can compute the estimates $\hat{\mu} = \bar{X}$ and $\hat{\nu}^2 = \frac{1}{N-1}\sum_{i=1}^N (X_i - \bar{X})^2$, and then to estimate $\bar{p}_X(x)$ by solving (numerically) the equation

$$x = \hat{\mu} + \frac{\hat{\nu}}{(1 - \alpha)\sqrt{2\pi}} e^{-z_\alpha^2/2}, \tag{3.2}$$

and setting the estimate $\hat{p}_X(x) = 1 - \hat{\alpha}$. For the estimates $\hat{\mu}$ and $\hat{\nu}$ their confidence intervals can be computed from the same sample. Consequently these confidence intervals can be used to construct a confidence interval for $\bar{p}_X(x)$.

Suppose now that the probability distribution of X is contaminated by another distribution. That is, for some $\gamma \in (0,1)$ and cdfs F_1 and F_2, the cdf F of X is given as convex combination $F(\cdot) = \gamma F_1(\cdot) + (1-\gamma)F_2(\cdot)$. We have then

$$\begin{aligned}
\mathsf{CVaR}_\alpha(F) &= \inf_{t\in\mathbb{R}} \left\{ t + (1-\alpha)^{-1}\mathbb{E}_F[X-t]_+ \right\} \\
&= \inf_{t\in\mathbb{R}} \left\{ t + (1-\alpha)^{-1}\gamma\mathbb{E}_{F_1}[X-t]_+ + (1-\alpha)^{-1}(1-\gamma)\mathbb{E}_{F_2}[X-t]_+ \right\} \\
&\geq \inf_{t_1,t_2\in\mathbb{R}} \left\{ \gamma t_1 + (1-\gamma)t_2 + (1-\alpha)^{-1}\gamma\mathbb{E}_{F_1}[X-t]_+ + (1-\alpha)^{-1}(1-\gamma)\mathbb{E}_{F_2}[X-t]_+ \right\} \\
&= \gamma \inf_{t_1\in\mathbb{R}} \left\{ t_1 + (1-\alpha)^{-1}\mathbb{E}_{F_1}[X-t]_+ \right\} + (1-\gamma) \inf_{t_2\in\mathbb{R}} \left\{ t_2 + (1-\alpha)^{-1}\mathbb{E}_{F_2}[X-t]_+ \right\} \\
&= \gamma\mathsf{CVaR}_\alpha(F_1) + (1-\gamma)\mathsf{CVaR}_\alpha(F_2).
\end{aligned}$$

That is

$$\mathsf{CVaR}_\alpha(F) \geq \gamma\mathsf{CVaR}_\alpha(F_1) + (1-\gamma)\mathsf{CVaR}_\alpha(F_2). \tag{3.3}$$

It follows that solution of equation (2.2) is smaller than solution of equation

$$x = \gamma\mathsf{CVaR}_\alpha(F_1) + (1-\gamma)\mathsf{CVaR}_\alpha(F_2). \tag{3.4}$$

Hence by computing solution $\tilde{\alpha}$ of equation (3.4) we obtain $\bar{p}_X(x) \geq 1 - \tilde{\alpha}$, i.e., $1 - \tilde{\alpha}$ gives a lower bound for $\bar{p}_X(x)$. In particular, if F_1 and F_2 are respective normal distributions $\mathcal{N}(\mu_1, \nu_1^2)$ and $\mathcal{N}(\mu_2, \nu_2^2)$, then equation (3.4) can be solved numerically using formula (3.1).

We can also approach estimation of $\bar{p}_X(x)$ by using the importance sampling method. That is, suppose that we have a reasonable estimate $F(\cdot)$ of the distribution of X, with respective density $f(\cdot) = F'(\cdot)$. Consider the transformation $Y^j := X^j + c$, $i = 1, ..., N$, of the sample, where c is a chosen constant. Then $\mathsf{CVaR}_\alpha(X)$ can be estimated by

$$\widehat{\mathsf{CVaR}}^c_{\alpha,N} := \inf_{t\in\mathbb{R}} \left\{ t + (1-\hat{\alpha})^{-1}N^{-1}\sum_{j=1}^N L(Y^j)\big[Y^j - t\big]_+ \right\}, \tag{3.5}$$

where $L(y) := f(y)/f(y-c)$ is the corresponding likelihood ratio. By computing solution $\tilde{\alpha}_N$ of equation

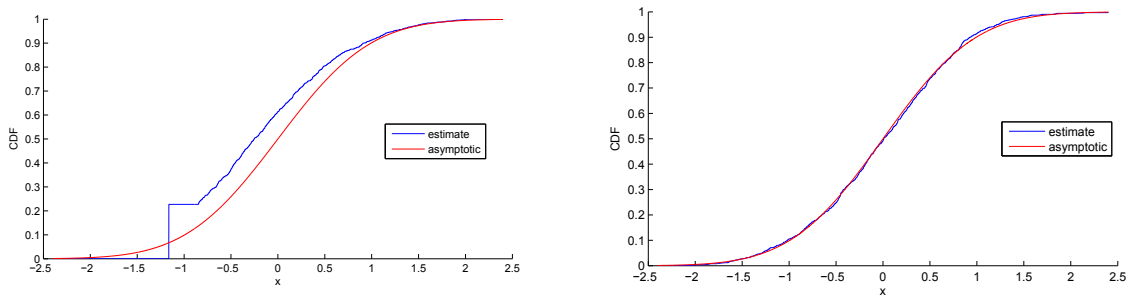
$$x = \widehat{\mathsf{CVaR}}^c_{\alpha,N},$$

we obtain an estimate $1 - \tilde{\alpha}_N$ of $\bar{p}_X(x)$. Intuitively "good choice" of constant c should be such that $Y^j - t^*$ is positive for many of Y^j values (here t^* is the minimizer in the right hand side of (3.5)). This would require a further investigation.

4 Numerical illustration for exponentially distributed variables

Consider an exponentially distributed random variable X with distribution parameter $\lambda > 0$, and pdf $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and $f(x) = 0$ for $x < 0$. Note that $\mathbb{E}[X] = \lambda^{-1}$. The respective quantile function is $q_\alpha(X) = -\lambda^{-1}\ln(1-\alpha)$, and the Conditional Value-at-Risk is

$$\bar{q}_\alpha(X) = -\lambda^{-1}\big[\ln(1-\alpha) - 1\big],\ \alpha \in [0,1).$$



(a) $N = 10$ (b) $N = 500$

Figure 1: Asymptotic CDF and empirical CDF of $N^{1/2}\left(\hat{p}_N(x) - \bar{p}_x(X)\right)$ for an estimator $\hat{p}_N(x)$ for $x = 2$.

It is easy to see that the solution of equation $-\lambda^{-1}[\ln(1-\alpha) - 1] = x$ yields

$$\bar{p}_x(X) = 1 - \alpha = e^{-\lambda(x-\lambda^{-1})} = 1 - F_X(x - \lambda^{-1}), \text{ for } x > \mathbb{E}[X],$$

while for $x \leq \mathbb{E}[X]$ the bPOE value is $\bar{p}_x(X) = 1$. For the further calculations let us assume that $x \geq \mathbb{E}[X]$. Denote $\alpha = 1 - \bar{p}_x(X)$, then

$$\bar{a} = \frac{1}{x - q_\alpha(X)} = \frac{1}{x + \lambda^{-1}(-\lambda x + 1)} = \lambda,$$

i.e., the optimal value of a is independent of x for exponential distributions. Let us calculate the asymptotic variance of the bPOE estimator $\hat{p}_N(x)$. Since $\bar{a} = \lambda$ we have that asymptotic variance $\sigma^2(x)$, given in (2.14), can be written as

$$\sigma^2(x) = \mathbb{E}\big([\lambda(X - x) + 1]_+\big)^2 - \bar{p}_x(X)^2.$$

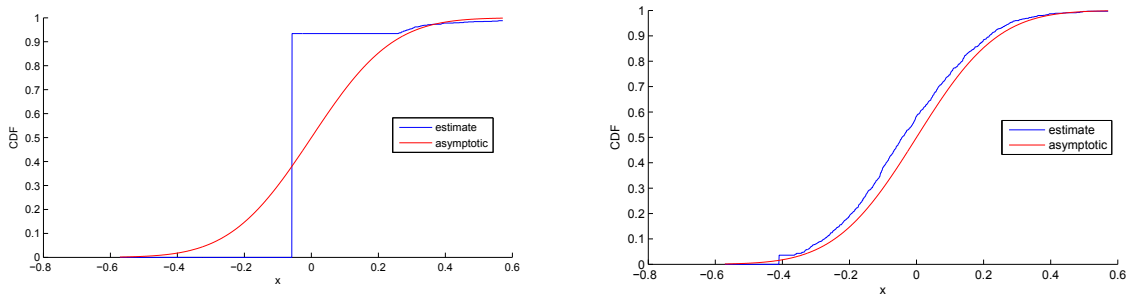
Denote $b := -\lambda x + 1$, then $\bar{p}_x(X) = e^b$, and

$$\sigma^2(x) = \int_{-b/\lambda}^{\infty} (\lambda\xi + b)^2 f(\xi) d\xi - e^{2b} = e^b \int_0^{\infty} \zeta^2 e^{-\zeta} d\zeta - e^{2b} = e^{-\lambda x + 1}(2 - e^{-\lambda x + 1}).$$

Let us illustrate Theorem 2.1. Namely, that $N^{1/2}\left(\hat{p}_N(x) - \bar{p}_x(X)\right)$ converges in distribution to normal $\mathcal{N}(0, \sigma^2(x))$. For $\lambda = 1$ and $x = 2$ the bPOE value is $\bar{p}_2(X) \approx 0.368$, while the asymptotic variance is $\sigma^2(2) \approx 0.6$. The weak convergence of $N^{1/2}\left(\hat{p}_N(2) - 0.368\right)$ to the asymptotic distribution $\mathcal{N}(0, 0.6)$ is illustrated in Figure 1. For $x = 5$ the bPOE value is $\bar{p}_5(X) \approx 0.018$ and the asymptotic variance is $\sigma^2(5) \approx 0.036$. It is expected that convergence will be slower for the larger x, see Figure 2 illustrating $N^{1/2}\left(\hat{p}_N(5) - 0.018\right)$ converges in distribution to $\mathcal{N}(0, 0.036)$. Similarly to the case of large x, convergence to asymptotically normal distribution can be slower for the values x close to $\mathbb{E}[X]$, see Figure 3 for an illustration. Figures 1–3 show that for considered example and reasonably large N, theoretical asymptotics were in reasonable agreement with experimental data.

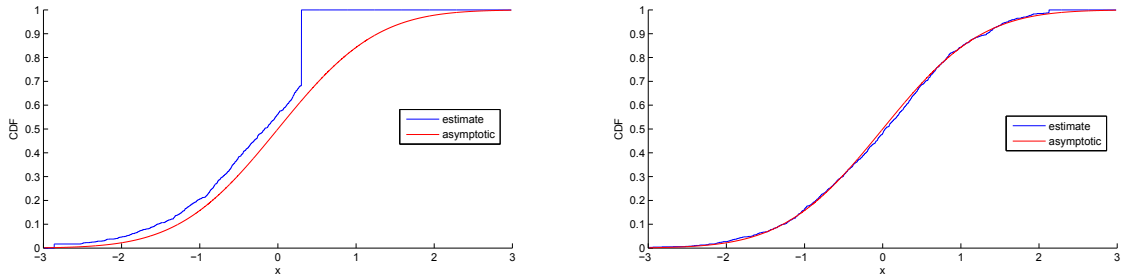
Let us illustrate bias estimate (2.21). Since $\gamma^2 = \mathbb{E}\left[\frac{\partial \Psi(\bar{a}, X)}{\partial a}\right]^2$ and

$$\frac{\partial \Psi(\bar{a}, X)}{\partial a} = \begin{cases} X - x, & \text{if } X \geq x - \frac{1}{a}; \\ 0, & \text{otherwise,} \end{cases},$$



(a) $N = 10$ (b) $N = 500$

Figure 2: Asymptotic CDF and empirical CDF of $N^{1/2}\left(\hat{p}_N(x) - \bar{p}_x(X)\right)$ for an estimator $\hat{p}_N(x)$ for $x = 5$.



(a) $N = 10$ (b) $N = 500$

Figure 3: Asymptotic CDF and empirical CDF of $N^{1/2}\left(\hat{p}_N(x) - \bar{p}_x(X)\right)$ for an estimator $\hat{p}_N(x)$ for $x = 1.1$.

we have

$$\gamma^2 = \int_{x-1/\lambda}^{\infty} (\xi - x)^2 \lambda e^{-\lambda\xi} d\xi = \frac{1}{\lambda^2} e^{-\lambda x} \int_{x-1/\lambda}^{\infty} (\lambda(\xi - x))^2 e^{-\lambda(\xi-x)} d\lambda(\xi - x) =$$
$$= \frac{1}{\lambda^2} e^{-\lambda x} \int_{-1}^{\infty} \zeta^2 e^{-\zeta} d\zeta = \frac{1}{\lambda^2} e^{-\lambda x+1}.$$

Note further that $\psi''(\bar{a}) = \frac{f(x-1/\bar{a})}{\bar{a}^3} = \frac{\lambda e^{-\lambda x+1}}{\lambda^3} = \gamma^2$. Therefore,

$$\mathbb{E}[\hat{p}_N(x)] - \bar{p}_x(X) = -\frac{1}{2N} + o(N^{-1}).$$

Remark 3 It could be noted that for larger N the theoretical asymptotic cdf is consistently below the corresponding estimated cdf. This behavior is consistent with the fact that the second order term (2.19), in the respective asymptotic expansion, is negative.

Now following the method of Section 3 aimed at rare events, denote $L := \sum_{j=1}^{N} L(Y^j)$ and rewrite (3.5) as

$$x = \inf_{t \in \mathbb{R}} \left\{ t + \frac{N^{-1}L}{(1-\hat{\alpha})} \sum_{j=1}^{N} \frac{L(Y^j)}{L} [Y^j - t]_+ \right\}.$$

Denote by $\tilde{p}_N^c(x)$ the importance sampling estimate, namely, $1 - \hat{\alpha}$. If $\sum_{j=1}^{N} \frac{L(Y^j)}{L} Y^j < x < \max_j Y^j$, then, following bPOE calculation formula,

$$\tilde{p}_N^c(x) = N^{-1} L \min_{a \geq 0} \sum_{j=1}^{N} \frac{L(Y^j)}{L} [a(Y^j - x) + 1]_+ = N^{-1} L \hat{p}_{N|L(X+c)}(x - c),$$

where $N|L(X+c)$ implies that the empirical distribution has probabilities $L(X^j + c)/L$ instead of equal probabilities N^{-1}. That is, the importance sampling estimate equals to a scaled regular estimate calculated under modified sample weights.

Note further that for an exponential variable the value $f(Y)/f(Y-c) = e^{-\lambda c}$, therefore, new weights are still uniform, and

$$\tilde{p}_N^c(x) = e^{-\lambda c} \hat{p}_N(x - c).$$

If $\hat{p}_N(x-c)$ estimates $\bar{p}_{x-c}(X) = e^{-\lambda(x-c)+1}$, then $\tilde{p}_N^c(x)$ estimates $e^{-\lambda c} e^{-\lambda(x-c)+1} = e^{-\lambda x+1} = \bar{p}_x(X)$, the true bPOE value. Note that this will only work when $x - c > \mathbb{E}[X]$, that is, $c < x - \frac{1}{\lambda}$. The variable $N^{1/2}(\tilde{p}_N^c(x) - \bar{p}_x(X)) = e^{-\lambda c} N^{1/2}(\hat{p}_N(x-c) - \bar{p}_{x-c}(X))$, therefore, converges in distribution to $\mathcal{N}(0, \sigma_c^2(x))$, where

$$\sigma_c^2(x) = (e^{-\lambda c})^2 \sigma^2(x-c) = 2e^{-2\lambda c - \lambda(x-c)+1} - e^{-2\lambda c - 2\lambda(x-c)+2} = 2e^{-\lambda(x+c)+1} - e^{2(-\lambda x+1)},$$

i.e., the variance decreases with increase of c and is minimal when $c \to x - \frac{1}{\lambda} \equiv c^*$, then $\sigma_c^2(x) \to (e^{-\lambda x+1})^2$ as opposed to $N^{1/2}(\hat{p}_N(x) - \bar{p}_x(X))$ converging to $\mathcal{N}(0, \sigma^2(x))$ with $\sigma^2(x) = e^{-\lambda x+1}(2 - e^{-\lambda x+1})$. The larger x, the smaller the ratio

$$\frac{\sigma_{c^*}^2(x)}{\sigma^2(x)} = \frac{e^{-\lambda x+1}}{2 - e^{-\lambda x+1}}.$$

5 Optimization of bPOE

Consider the following optimization problem

$$\min_{y\in\mathcal{Y}} \bar{p}_x(G(y,\xi)). \tag{5.1}$$

Here $\mathcal{Y}$ is a nonempty closed subset of $\mathbb{R}^n$, $\xi \in \mathbb{R}^d$ is a random vector and $G : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ is a *Carathéodory function*, i.e., $G(y,\cdot)$ is measurable for every y and $G(\cdot,\xi)$ is continuous for a.e. ξ (cf., [10, Example 14.29]). We assume that $\mathbb{E}|G(y,\xi)| < \infty$ for every $y \in \mathcal{Y}$, and hence the function $\bar{p}_x(G(y,\xi))$ is well defined and finite valued on $\mathcal{Y}$. We also assume that the set $\mathcal{Y}$ is *convex* and $G(\cdot,\xi)$ is *convex* for a.e. ξ. In particular we consider piecewise affine functions of the form[6]

$$G(y,\xi) := \max_{1\le i\le m} \big\langle b_i(\xi), y\big\rangle + c_i(\xi), \tag{5.2}$$

with $b_i : \mathbb{R}^d \to \mathbb{R}^n$ and $c_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1,...,m$, being measurable. Note that by (2.4) we have that if $x \ge G(\bar{y},\xi)$ for some $\bar{y} \in \mathcal{Y}$ and a.e. ξ, then $\bar{p}_x(G(\bar{y},\xi)) = 0$ and hence $\bar{y}$ is an optimal solution of problem (5.1).

By (2.5) we can write problem (5.1) in the form

$$\min_{a\ge 0,\, y\in\mathcal{Y}} \psi(a,y), \tag{5.3}$$

where $\Psi(a,y,\xi) := [a(G(y,\xi) - x) + 1]^+$ and $\psi(a,y) := \mathbb{E}[\Psi(a,y,\xi)]$. The Sample Average Approximation (SAA) of problem (5.3) is the problem

$$\min_{a\ge 0,\, y\in\mathcal{Y}} \hat{\psi}_N(a,y), \tag{5.4}$$

where $\hat{\psi}_N(a,y) := N^{-1}\sum_{j=1}^N \Psi(a,y,\xi^j)$ with $\xi^1,...,\xi^N$ being an iid sample of the random vector ξ. That is, the SAA problem is obtained by replacing the probability distribution of ξ with its empirical estimate based on the generated random sample.

Let us consider the following reformulation of problems (5.3) and (5.4) (cf., [7]). By Fenchel-Moreau Theorem we have

$$G(y,\xi) = \sup_{y^*\in\mathbb{R}^n} \langle y^*, y\rangle - G^*(y^*,\xi), \tag{5.5}$$

where $G^*(y^*,\xi)$ is the conjugate of $G(y,\xi)$ given by

$$G^*(y^*,\xi) = \sup_{y\in\mathbb{R}^n} \langle y^*, y\rangle - G(y,\xi). \tag{5.6}$$

Note that $G^*(y^*,\xi)$ can take value $+\infty$ and the maximization in (5.5) is performed over domain

$$D(\xi) := \{y^* \in \mathbb{R}^n : G^*(y^*,\xi) < +\infty\} \tag{5.7}$$

of $G^*(\cdot,\xi)$.

By replacing $y \in \mathbb{R}^n$ with $(y,1) \in \mathbb{R}^{n+1}$ in (5.5), we can write

$$G(y,\xi) - x = \sup_{y^*\in D(\xi)} \big\langle \big(y^*, -G^*(y^*,\xi) - x\big), (y,1)\big\rangle. \tag{5.8}$$

[6] By $\langle x,y\rangle$ we denote the standard scalar product of vectors $x, y \in \mathbb{R}^n$.

Consequently we can formulate problem (5.3) in the following equivalent form

$$\min_{z\in\mathcal{Z}} \mathbb{E}[\bar{G}(z,\xi)+1]^+, \tag{5.9}$$

where

$$\mathcal{Z} := \big\{z\in\mathbb{R}^{n+1} : z=(ay,a),\ y\in\mathcal{Y},\ a\geq 0\big\} \tag{5.10}$$

and

$$\bar{G}(z,\xi) := \sup_{y^*\in D(\xi)} \big\langle \big(y^*, -G^*(y^*,\xi)-x\big), z\big\rangle . \tag{5.11}$$

In particular, if $G(y,\xi)$ is of the form (5.2), then

$$\bar{G}(z,\xi) = \max_{1\leq i\leq m} \big\langle \big(b_i(\xi), c_i(\xi)-x\big), z\big)\big\rangle . \tag{5.12}$$

The corresponding SAA problem is

$$\min_{z\in\mathcal{Z}} \frac{1}{N}\sum_{j=1}^{N} \big[\bar{G}(z,\xi^j)+1\big]_+ . \tag{5.13}$$

Convexity of $\mathcal{Y}$ implies that the set (cone) $\mathcal{Z}$ is convex. If the set $\mathcal{Y}$ is polyhedral, defined by a finite number of linear constraints, then the cone $\mathcal{Z}$ is also polyhedral. If, moreover, $G(x,\xi)$ is of the form (5.2), and hence $\bar{G}(x,\xi)$ is of the form (5.12), then the SAA problem (5.13) can be written as a linear programming problem (cf., [7]). In general closedness of the set $\mathcal{Y}$ does not imply that the cone $\mathcal{Z}$ is closed. The cone $\mathcal{Z}$ is closed in two important cases, namely when $\mathcal{Y}$ is polyhedral or compact. Anyway by continuity arguments the optimal values of problems (5.9) and (5.13) are not changed if the set $\mathcal{Z}$ is replaced by its topological closure $\mathrm{cl}(\mathcal{Z})$.

Denote by ϑ^* and $\hat{\vartheta}_N$ the optimal values of problems (5.9) and (5.13), respectively. Note again that ϑ^* is the optimal value of the "true" problem (5.1) and $\hat{\vartheta}_N$ is the optimal value of its SAA counterpart. Also let $\mathcal{Z}^*$ be the set of optimal solutions of problem (5.9), and $\hat{\mathcal{Z}}_N$ be the set of optimal solutions of the SAA problem (5.13) with $\mathcal{Z}$ replaced by its closure $\mathrm{cl}(\mathcal{Z})$. It could be noted that $(\bar{a},\bar{y})$ is an optimal solution of problem (5.3) iff $\bar{z}=\bar{a}(\bar{y},1)$ is an optimal solution of problem (5.9), and similarly for the corresponding SAA problems. Hence if the cone $\mathcal{Z}$ is closed, then the set $\mathcal{Z}^*$ consists of points $\bar{z}=\bar{a}(\bar{y},1)$ with $\bar{a},\bar{y}$ being an optimal solution of problem (5.3).

Denote $\mathbb{D}(A,B) := \sup_{y\in A}\mathrm{dist}(y,B)$ the deviation of set $A\subset\mathbb{R}^n$ from set $B\subset\mathbb{R}^n$.

Theorem 5.1 *Suppose that the optimal set $\mathcal{Z}^*$ is nonempty and bounded. Then $\hat{\vartheta}_N$ converges w.p.1 to ϑ^* and $\mathbb{D}(\hat{\mathcal{Z}}_N,\mathcal{Z}^*)$ converges w.p.1 to 0 as $N\to\infty$.*

Proof The function $[\bar{G}(\cdot,\xi)+1]_+$ is convex and the set $\mathrm{cl}(\mathcal{Z})$ is convex. Therefore we can apply [13, Theorem 5.4] to conclude the proof. ∎

By using [13, Theorem 5.7], together with representation (5.9), we can write the following asymptotics of the SAA estimator $\hat{\vartheta}_N$.

Theorem 5.2 *Suppose that:* (i) *the optimal set $\mathcal{Z}^*=\{\bar{z}\}$ is a singleton,* (ii) *variance*

$$\sigma^2 = \mathrm{Var}\big([\bar{G}(\bar{z},\xi)+1]_+\big)$$

is finite, (iii) *there exists a measurable function $C(\xi)$ such that $\mathbb{E}[C(\xi)^2]<\infty$ and*

$$|\bar{G}(z,\xi)-\bar{G}(z',\xi)|\leq C(\xi)\|z-z'\|$$

for all $z,z'\in\mathcal{Z}$ and a.e. ξ.

Then $N^{1/2}(\hat{\vartheta}_N-\vartheta^*)$ converges in distribution to normal $\mathcal{N}(0,\sigma^2)$.

6 Minimum Volume Ellipsoid Problem: modification with bPOE

The *Minimum Volume Ellipsoid* (MVE) problem is a problem of covering K out of N data points $\xi^1, ..., \xi^N$ with an ellipsoid while minimizing the volume of such ellipsoid. If $K = N$, then the problem is called minimum *covering* ellipsoid problem, since all data points must lie inside the ellipsoid. This problem can be formulated as a convex problem and solved very efficiently [15]. The general MVE problem, with $K < N$, is a hard non-convex problem. A lot of efforts were directed towards finding an approximate solutions, or towards finding good heuristics to solve a similar problem and achieve a similar solution.

Minimum volume ellipsoid problems were studied extensively from perspectives of optimization [3, 6, 15], statistics [16, 4], and machine learning [5, 1, 14, 17, 2]. Current results include various optimization algorithms and applications, primarily in machine learning, as well as statistical properties of the MVE estimator. The general MVE is known for its high resistance to outliers and high (up to 0.5) breakdown value [4]. Below we present a new bPOE-MVE estimator. Similar to covering-MVE estimator, it can be computed efficiently as a convex optimization problem. Unlike covering-MVE and similar to MVE, the bPOE-MVE allows data points to lie outside the ellipsoid, but, unlike MVE, it accounts for actual positions of data points lying outside. This is why bPOE-MVE might not be as good as MVE in identifying outliers. Rather than that, we position this new estimator as a good tool to treat "problematic" rare points, which, even if may seem like outliers, appear in datasets on a regular basis.

Suppose that ξ is a random vector of dimension n. Let us parameterize an ellipsoid in $\mathbb{R}^n$ with its center $c \in \mathbb{R}^n$ and its positive definite shape matrix[7] $Q \in \mathbb{S}^{n\times n}_{++}$, such that the set $\{z \in \mathbb{R}^n : (z-c)^\mathsf{T} Q(z-c) = 1\}$ corresponds to the surface of the ellipsoid. Then covering the distribution ξ with an ellipsoid means satisfying the inequality $(\xi - c)^\mathsf{T} Q(\xi - c) \leq 1$ almost surely. Note that this problem has a solution only when the support $\Xi \subset \mathbb{R}^n$, of the distribution of random vector ξ, is bounded.

In order to make the corresponding problem convex let us introduce the following change of variables, $A := Q^{1/2}$ and $b := Ac$. Then, $(\xi - c)^\mathsf{T} Q(\xi - c) = \|A\xi - b\|_2$. Volume of the ellipsoid, parameterized by Q, equals $V = \det(Q^{-1})$ and minimization of ellipsoid volume is equivalent to maximization of $V^{-1/2} = \det(A)$, or to minimization of $-V^{-1/2n} = -(\det A)^{1/n}$, which is a convex function of A[8]. Note that ellipsoid covering of points $\xi^1, \ldots, \xi^N$ means that $\|A\xi^i - b\|_2 \leq 1$ for $1 \leq i \leq N$, and the covering-MVE problem is reduced to a finite-dimensional convex programming problem. General MVE problem for the random vector ξ can be written as the following chance constrained problem

$$\begin{array}{cl} \min\limits_{A\in\mathbb{S}^{n\times n}_{++}, b\in\mathbb{R}^n} & -(\det A)^{1/n} \\ \text{s.t.} & \mathsf{Prob}\big(\|A\xi - b\|_2 \geq 1\big) \leq \alpha. \end{array} \tag{6.14}$$

Note that if random vector ξ has finite support $\Xi = \{\xi^1, \ldots, \xi^N\}$, with assigned equal probabilities, and $\alpha = (N-K)/N$, then problem (6.14) corresponds to the problem of covering at least K points from the set $\{\xi^1, \ldots, \xi^N\}$ with an ellipsoid, while minimizing volume of this ellipsoid. If $\alpha = 0$, then MVE problem becomes covering-MVE and convex. The considered

[7] By $\mathbb{S}^{n\times n}$ we denote the linear space of $n \times n$ symmetric matrices, and by $\mathbb{S}^{n\times n}_{++}$ its subset of positive definite matrices.

[8] To minimize volume, we minimize a convex function $-(\det A)^{1/n}$ rather than a commonly used and also convex function $-\ln\det A$, mostly for the convenience of numerical experiments, since our optimization program code has started from `http://cvxr.com/cvx/examples/cvxbook/Ch08_geometric_probs/html/min_vol_elp_finite_set.html`.

MVE problem has two parameters, the ellipsoid volume V and a measure of the area outside the ellipsoid parameterized by α. The problem (6.14) constraints a measure outside of ellipsoid with parameter α and minimizes volume. It is clear that if value of parameter α increases from 0 to 1, then the optimal volume decreases from $+\infty$ to 0. An alternative problem statement would require the volume to be no greater than V, and minimize the measure of the outside area. That is, another view of the problem (6.14) is obtained when the objective and the constraint switch places:

$$\begin{array}{ll} \min\limits_{A\succ 0,b} & p_1(\|A\xi - b\|_2) \\ \text{s.t.} & (\det A)^{1/n} \geq V^{-1/2n}, \end{array} \tag{6.15}$$

where

$$p_1(\|A\xi - b\|_2) := \mathsf{Prob}\big(\|A\xi - b\|_2 \geq 1\big).$$

It is easy to see that the two parametric problem families with parameters α and V share the same frontier of optimal solutions. Consider set $S = \{(V, \alpha) : V = \det A^{-1/2}, \alpha = p_1(\|A\xi - b\|_2), A \succ 0\}$, and suppose that a certain pair $(V_0, \alpha_0) \in S$ is dominant, i.e. $S \cap [0, V_0] \times [0, \alpha_0] = (V_0, \alpha_0)$. Then problem (6.14) with parameter α_0 and problem (6.15) with parameter V_0 will have the same optimal solutions. Furthermore, the probability of exceedance $p_1(\|A\xi - b\|_2)$ can be changed to the bPOE $\bar{p}_1(\|A\xi - b\|_2)$, which is the smallest quasi-convex upper bound for POE, and allows for the following formulation:

$$\begin{array}{ll} \min\limits_{A\succ 0,b} & \bar{p}_1(\|A\xi - b\|_2) \\ \text{s.t.} & (\det A)^{1/n} \geq V^{-1/2n}. \end{array} \tag{6.16}$$

This problem, as the previous one, has an upper bound on an ellipsoid volume and minimizes a measure of the furthest from the ellipsoid center points such that "on average" these points lie on the surface, i.e., conditionally expected value of $\|A\xi - b\|_2$ for these points is 1.

An alternative way to convexify the chance constrained problem (6.14) is to substitute the constraint with $\mathrm{CVaR}_\alpha(\|A\xi - b\|_2) \leq 1$. Note here that CVaR constraint is equivalent to $\bar{p}_x(\|A\xi - b\|_2) \leq \alpha$. Therefore, the correspondence between CVaR-MVE problem family and (6.16) bPOE-MVE problem family is exactly the same as the one between (6.14) and (6.15). The formulation (6.16) allows for a convex reformulation, as shown in Section 5. It is easy to see that the function $G(y, \xi) := \|A\xi - b\|_2$, where $y = (A, b) \in \mathcal{Y} := \{(A, b) : A \in \mathbb{S}^{n\times n}_{++}, b \in \mathbb{R}^n\}$, is a Caratheodory function. Since $G(\cdot, \xi)$ is positive homogeneous, i.e., $G(\lambda y, \xi) = \lambda G(y, \xi)$ for $\lambda \geq 0$, its convex conjugate $G^*(y^*, \xi)$, defined in (5.6), is the indicator function of its domain $D(\xi)$ (see (5.7) for definition of $D(\xi)$). Rewriting (5.10) we get

$$\begin{aligned} \mathcal{Z} &= \{z = (ay, a) : y \in \mathcal{Y}, a \geq 0\} = \\ &= \{z = (aA, ab, a) : A \in \mathbb{S}^{n\times n}_{++}, b \in \mathbb{R}^n, a \geq 0\} = \\ &= \{z = (B, d, a) : B \in \mathbb{S}^{n\times n}_{++}, d \in \mathbb{R}^n, a \geq 0\}. \end{aligned}$$

When optimal z is obtained, optimal solution to the original problem (6.16) can be obtained from $z = (B, d, a)$ and equations $B = aA$ and $d = ab$. Using $G^*(y^*, \xi) = 0$ for $y^* \in D(\xi)$, we rewrite (5.11):

$$\begin{aligned} \bar{G}(z, \xi) = \bar{G}((B, d, a), \xi) &= \sup_{y^* \in D(\xi)} \langle (y^*, -G^*(y^*, \xi) - 1), (B, d, a)\rangle = \\ &= -a + \sup_{y^* \in D(\xi)} \langle y^*, (B, d)\rangle - G^*(y^*, \xi) = \\ &= G((B, d), \xi) - a = \|B\xi - d\|_2 - a. \end{aligned}$$

Therefore, rewriting (5.13), we can write the corresponding SAA problem:

$$\begin{array}{ll} \min\limits_{B,d,a} & \sum_{i=1}^{N} \left[\|B\xi^i - d\|_2 - a + 1\right]_+ \\ \text{s.t.} & (\det B)^{1/n} \geq aV^{-1/2n}, \\ & B \succ 0,\ a \geq 0. \end{array} \tag{6.17}$$

The MVE problems are suited perfectly for multidimensional distributions coming from the elliptical class, that is, when probability density function (pdf) has the form $f(x) = g\big(\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}\big)$, where g is a one-dimensional density function. For bPOE-MVE problem, we show below that the solution for arbitrary Σ and μ may be obtained from the solution to the problem with identity shape matrix $\Sigma = I$ and zero mean $\mu = 0$. Suppose that random vector ξ has pdf $g\big(\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}\big)$, and the function g is decreasing on $[0, +\infty)$, then $\nu := \Sigma^{-1/2}(\xi - \mu) \propto g(\sqrt{y^T y})$. Denote by B_Σ, d_Σ, a_Σ the optimal solution to the problem for the original random vector ξ with volume constraint parameter V_Σ:

$$\begin{array}{ll} \min\limits_{B,d,a} & \mathbb{E}\big[\|B\xi - d\|_2 - a + 1\big]_+ \\ \text{s.t.} & (\det B)^{1/n} \geq aV_\Sigma^{-1/2n}, \\ & B \succ 0,\ a \geq 0. \end{array} \tag{6.18}$$

For the "standardized" random vector ν, take parameter value V_I and denote the optimal solution by B_I, d_I, a_I, and apply $\nu = \Sigma^{-1/2}(\xi - \mu)$:

$$\begin{array}{ll} \min\limits_{B,d,a} & \mathbb{E}[\|B\nu - d\|_2 - a + 1]_+ = \mathbb{E}[\|B\Sigma^{-1/2}\xi - (B\Sigma^{-1/2}\mu + d)\|_2 - a + 1]_+ \\ \text{s.t.} & (\det B)^{1/n} \geq aV_I^{-1/2n} \Leftrightarrow \left(\det(B\Sigma^{-1/2})\right)^{1/n} \geq a(\det\Sigma)^{-1/2n}V_I^{-1/2n}, \\ & B \succ 0,\ a \geq 0. \end{array}$$

Assume that $V_\Sigma = V_I \det\Sigma$, then optimal solutions are connected as follows: $B_\Sigma = B_I\Sigma^{-1/2}$, $d_\Sigma = d_I + B_\Sigma\mu$, $a_\Sigma = a_I$. Therefore, for our purposes, it can be assumed without loss of generality that the elliptical distribution is "standard": it has a zero mean and an identity shape matrix. Further on, because of a symmetry of such distribution, which is a spherical distribution, the optimal ellipsoid is a sphere with a center at zero. Hence $d_I = 0$ and $B_I = a_I V_I^{-1/2n} I$. Noting that $B = aA$ and $d = ab$, we get that in the original problem $A_I = V_I^{-1/2n} I$ and $b_I = 0$. Note also that the same calculations are valid for probability of exceedance minimization with a volume constraint. Therefore, for a class of elliptical distributions, POE minimization and bPOE minimization provide the same optimal solution, but different objective values.

To test the new covering ellipsoid problem behavior, we are varying the tail fatness. One of the ways to do that is to consider a function g from exponential power distribution, $g(x) = C \cdot e^{-\frac{|x|^\beta}{\alpha}}$, see Figure 4 for an illustration. Normal distribution is a special case of the exponential power distribution with $\beta = 2$.

We generate samples from elliptical exponential power distribution in a following way. We generate a uniformly distributed on a unit sphere random vector $U = X/\|X\|_2$, where $X \propto \mathcal{N}(0, I)$ is a standard normal vector. We generate variable $R \geq 0$ from radial distribution, such that variable UR has a density proportional to $g(\|x\|_2) = C \cdot e^{-\|x\|_2^\beta/\alpha}$. Hence, density for R must be proportional to $x^{n-1}e^{-x^\beta/\alpha}$, therefore, CDF for R is proportional to

$$\int_0^x \zeta^{n-1}e^{-\zeta^\beta/\alpha}d\zeta = C \cdot \int_0^{x^\beta/\alpha} t^{n/\beta-1}e^{-t}dt = C \cdot \gamma(n/\beta, x^\beta/\alpha),$$

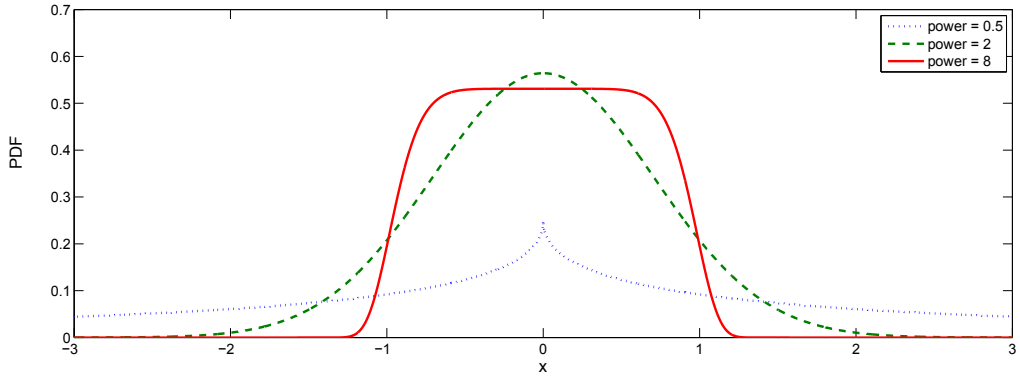


Figure 4: Probability density functions of exponential power distributions for power $\beta \in \{0.5, 2, 8\}$ and scale $\alpha = 1$.

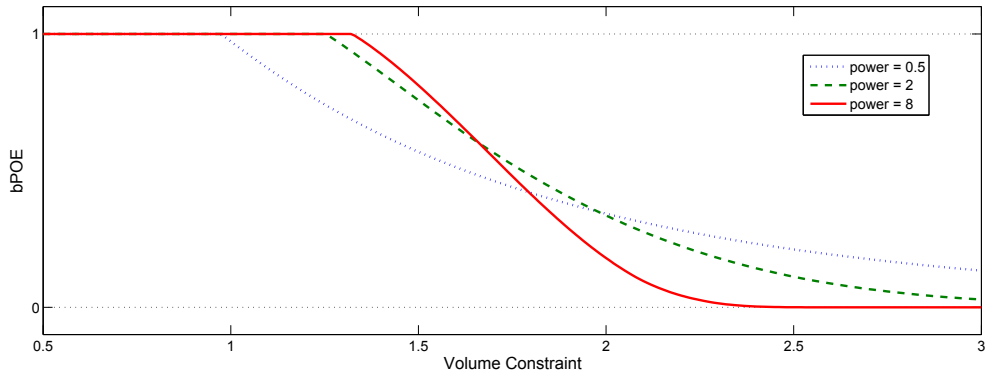


Figure 5: Optimal bPOE value ϑ^* (vertical axis) for the problem (6.18) as a function of the volume constraint parameter V (horizontal axis), for elliptical exponential power distribution with power values $\beta \in \{0.5, 2, 8\}$ and scale values α such that all covariation matrices are equal to identity matrix.

where $\gamma(a, x) \equiv \int_0^x t^{a-1} e^{-t} dt$ is an incomplete Gamma-function. Note that CDF of Gamma distribution with parameters k, θ is $C \cdot \gamma(k, \theta x)$. Therefore, if $G \propto \text{Gamma}(n/\beta, 1)$ and $R = \alpha Y^{1/\beta}$, then R has the required distribution. Finally, $\xi = RU\Sigma^{-1/2} + \mu$ has the requested elliptical exponential power distribution.

To compare how the optimal bPOE value decreases for distributions with different power, we fit scaling factors to make covariance matrices equal to identity matrix. Then we vary the upper bound V for the ellipsoid volume and measure optimal bPOE, see Figure 5 for an illustration. It can be seen that the optimal objective value ϑ^* decreases slower for the distributions with heavier tails.

Further on, we will compare convergence of solutions to the true optimum, as Theorem 5.1 predicts. Since convergence to the true optimal solution and to the asymptotic distribution is slower for both large and small values of bPOE, we take such values of volume constraints for different power values that true optimal bPOE values are equal to $1/2$. For the large enough sample size $N = 10^6$ we estimate the true optimal solution $\bar{z} = (\bar{B}, \bar{d}, \bar{a})$ to (6.18). First, we test convergence almost surely for optimal to (6.17) solutions $\hat{z}_N = (\hat{B}_N, \hat{d}_N, \hat{a}_N)$ by measuring

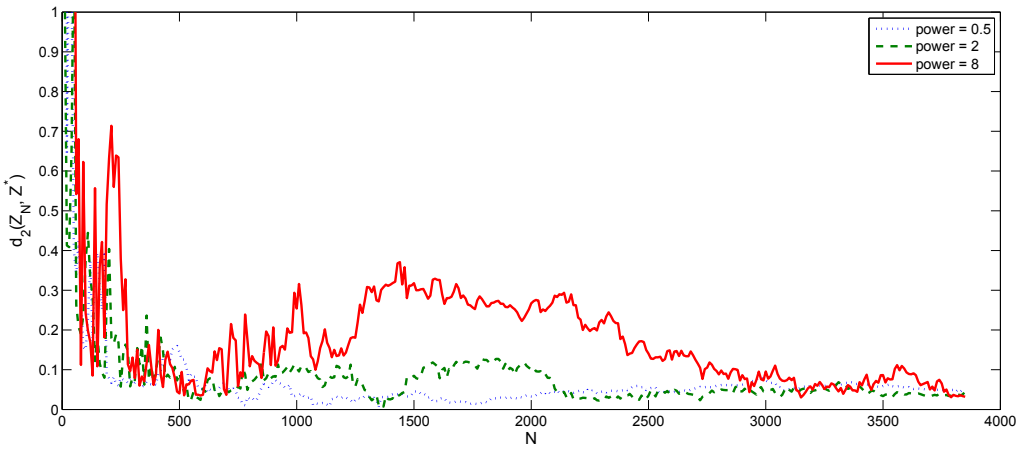


Figure 6: Euclidean norm of error $d_2(\hat{z}_N, \bar{z}) \equiv \|\hat{z}_N - \bar{z}\|_2$ for optimal to (6.17) solution $\hat{z}_N = (\hat{B}_N, \hat{d}_N, \hat{a}_N)$ by the sample size N for elliptical exponential power distribution with power $\beta \in \{0.5, 2, 8\}$. It can be seen that errors converge to 0 with N increasing for all power values.

$\|\bar{z} - \hat{z}_N\|_2$ and varying N, see Figure 6. We also measure error $\hat{\vartheta}_N - \vartheta^*$ of optimal values, where $\hat{\vartheta}_N$ is an optimal value to (6.17), and $\vartheta^* = 1/2$ is an optimal value to (6.18). See Figure 7 for an illustration. It can be noted that both optimal solutions and optimal values converge to the true ones, and that the fluctuations from the true optimum are higher for the lower power values, i.e., for distributions with heavier tails.

Theorem 5.2 shows that the scaled optimal objective values $N^{1/2}(\hat{\vartheta}_N - \vartheta^*)$ converge in distribution to the normal distribution $\mathcal{N}(0, \sigma^2)$, where $\sigma^2 = \mathrm{Var}([\bar{G}(\bar{z}, \xi) + 1]^+)$. With a large sample ($10^6$ observations) we get estimates for σ^2. For the values of $N = 50$ and $N = 500$ we generate $M = 1000$ samples of size N to estimate an empirical distribution of $\hat{\vartheta}_N$. We denote $\eta_N = N^{1/2}(\hat{\vartheta}_N - \vartheta^*)$, and by $\hat{\mathbb{E}}\eta_N$ and $\hat{\sigma}^2(\eta_N)$ we denote average and standard deviation of η_N among M generated samples. Table 1 contains experiment setup and measurements. Note that distributions with heavier tails have larger value of asymptotic variance, but smaller bias. That is, distributions can not be easily ranked on their convergence speed based on their tail heaviness. Note also that with N increasing both bias and estimated variance converge to theoretically predicted values, which supports the result of the theorem. While absolute values of variance are smaller for distributions with lighter tails, values of variance relative to corresponding asymptotic variance are approximately the same among the distributions. For empirical CDFs of η_{50}, η_{500}, and the asymptotic normal distribution, see Figure 8.

Acknowledgement

This work was partially supported by the USA AFOSR grants: "Design and Redesign of Engineering Systems", FA9550-12-1-0427, and "New Developments in Uncertainty: Linking Risk Management, Reliability, Statistics and Stochastic Optimization", FA9550-11-1-0258.

References

[1] Abou-Moustafa, K., and Ferrie, F. Regularized minimum volume ellipsoid metric for

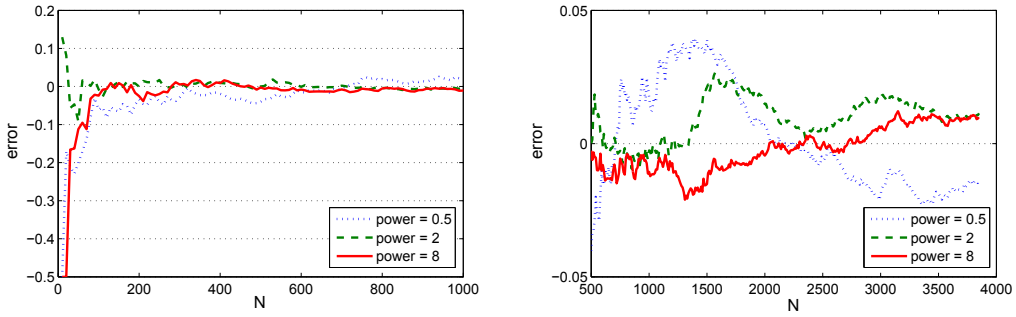
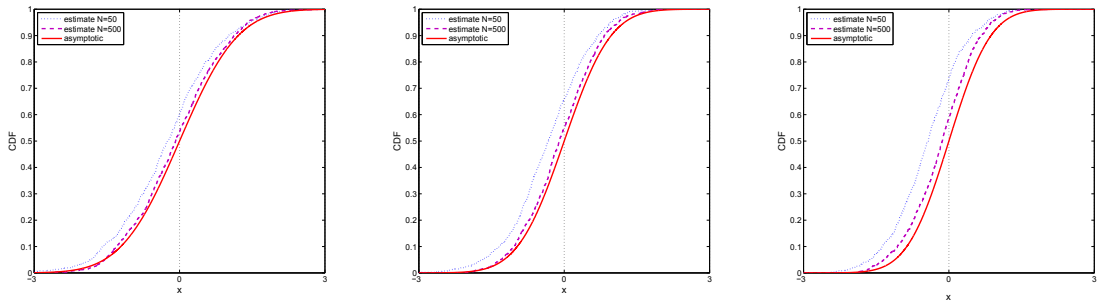


Figure 7: Error $\hat{\vartheta}_N - \vartheta^* = \hat{\vartheta}_N - 1/2$ of the optimal to (6.17) bPOE value $\hat{\vartheta}_N$ by the sample size N for elliptical exponential power distribution with $\beta \in \{0.5, 2, 8\}$. It can be noted, see left figure presenting interval $N \in [1, 1000]$, that errors converge to 0 with N increasing for all power values. Right figure presents interval $N \in [500, 4000]$. Note that the scale required to plot error values (< 0.05) for $N \in [500, 4000]$ is much smaller than the one required for $N \in [1, 1000]$, with error values ≤ 0.5.

power, β	scale, α	V	ϑ^*	σ^2	$\hat{\mathbb{E}}\eta_{50}$	$\hat{\sigma}^2(\eta_{50})$	$\hat{\mathbb{E}}\eta_{500}$	$\hat{\sigma}^2(\eta_{500})$
0.5	1	$1.2 \cdot 10^6$	0.5	0.9	-0.26	1	-0.07	0.83
2	1	2.5	0.5	0.57	-0.33	0.64	-0.09	0.54
8	1	0.54	0.5	0.46	-0.44	0.51	-0.17	0.44

Table 1: Setup of the experiment on weak convergence to the asymptotic normal distribution and obtained estimates. Values β, α are the power and the scale parameters of the power distribution; V is the upper volume constraint; ϑ^* is the optimal value of (6.18); σ^2 is the asymptotic variance for the estimator $\hat{\vartheta}_N$; $\eta_N = N^{1/2}(\hat{\vartheta}_N - \vartheta^*)$; $\hat{\mathbb{E}}$ is an average taken over $M = 1000$ generated samples; $\hat{\sigma}^2$ is an empirical variance calculated over M generated samples. Note that $0 > \hat{\mathbb{E}}\eta_N \to 0$ and $\hat{\sigma}^2(\eta_N) \to \sigma^2$ as $N \to \infty$, as predicted.



(a) $\beta = 0.5$ (b) $\beta = 2$ (c) $\beta = 8$

Figure 8: Asymptotic normal $\mathcal{N}(0, \sigma^2)$ CDF and empirical CDFs of $N^{1/2}\left(\hat{\vartheta}_N - \vartheta^*\right)$ for elliptical power distribution with power β.

query-based learning. In *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on* (2008), IEEE, pp. 188–193.

[2] Abou-Moustafa, K. T., and Ferrie, F. P. The minimum volume ellipsoid metric. In *Pattern Recognition*. Springer, 2007, pp. 335–344.

[3] Cook, R., Hawkins, D., and Weisberg, S. Exact iterative computation of the robust multivariate minimum volume ellipsoid estimator. *Statistics & probability letters 16*, 3 (1993), 213–218.

[4] Davies, L. The asymptotics of rousseeuw's minimum volume ellipsoid estimator. *The Annals of Statistics* (1992), 1828–1843.

[5] Hadi, A. S. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)* (1992), 761–771.

[6] Kumar, P., and Yildirim, E. A. Minimum-volume enclosing ellipsoids and core sets. *Journal of Optimization Theory and Applications 126*, 1 (2005), 1–21.

[7] Mafusalov, A., and Uryasev, S. Buffered probability of exceedance: Mathematical properties and optimization algorithms. Tech. rep., Research Report 2014-1, ISE Dept., University of Florida, 2014.

[8] Norton, M., and Uryasev, S. Maximization of auc and buffered auc in classification. Tech. rep., Research Report 2014-2, ISE Dept., University of Florida, 2014.

[9] Rockafellar, R. Safeguarding strategies in risky optimization. In *Presentation at the International Workshop on Engineering Risk Control and Optimization, Gainesville, FL* (2009).

[10] Rockafellar, R., and Wets, R. Variational analysis. 1998.

[11] Rockafellar, R. T., and Royset, J. O. On buffered failure probability in design and optimization of structures. *Reliability Engineering & System Safety 95*, 5 (2010), 499–510.

[12] Rockafellar, R. T., and Uryasev, S. Optimization of conditional value-at-risk. *Journal of risk 2* (2000), 21–42.

[13] Shapiro, A., Dentcheva, D., and Ruszczyński, A. *Lectures on stochastic programming: modeling and theory*, vol. 16. SIAM, 2014.

[14] Shivaswamy, P. K., and Jebara, T. Ellipsoidal machines. In *International Conference on Artificial Intelligence and Statistics* (2007), pp. 484–491.

[15] Sun, P., and Freund, R. M. Computation of minimum-volume covering ellipsoids. *Operations Research 52*, 5 (2004), 690–706.

[16] Van Aelst, S., and Rousseeuw, P. Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics 1*, 1 (2009), 71–82.

[17] Wei, X., Löfberg, J., Feng, Y., Li, Y., and Li, Y. Enclosing machine learning for class description. In *Advances in Neural Networks–ISNN 2007*. Springer, 2007, pp. 424–433.