European Journal of Operational Research 267 (2018) 1003–1013

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Expected shortfall: Heuristics and certificates

Federico Alessandro Ramponi*, Marco C. Campi

Università degli Studi di Brescia, Dipartimento di ingegneria dell'informazione, Via Branze 38, Brescia 25123, Italy

ARTICLE INFO

Article history:
Received 17 May 2016
Accepted 12 November 2017
Available online 20 November 2017

Keywords:
Stochastic programming
Convex programming
Scenario approach
Expected shortfall
CVaR

ABSTRACT

We consider the expected shortfall, a coherent risk measure that is gaining popularity outside mathematical finance and that is being applied to an increasing number of optimization problems due to its versatility and pleasant properties. A commonly used heuristic to optimize the expected shortfall consists in replacing the unknown distribution of the loss function with its empirical discrete counterpart constructed from observations. The boundary of the empirical shortfall tail is called the shortfall threshold, and, in this paper, we study the probability of incurring losses larger than the shortfall threshold. In a stationary set-up, we show that under mild conditions a striking universal result holds which says that the probability of losses exceeding the shortfall threshold is a random variable whose distribution is independent of the distribution of the loss function. This result complements previous findings on the expected shortfall and bears important practical consequences on the applications of this risk measure to stochastic optimization. The theory this result rests on is fully developed in this paper and its use is illustrated by examples.

1. Introduction

In this paper, we consider the following sample-based optimization problem:

$$x_N^* = \arg\min_{x\in\mathcal{X}} \{\text{average of the } k \text{ largest values among } L(x,\delta_1),\dots,L(x,\delta_N)\}, \quad (1)$$

where k is an integer in the range $1 \le k \le N$, $\mathcal{X}$ is a convex subset of $\mathbb{R}^d$, $L(\cdot,\delta_i)$ are convex cost functions, each one depending on a value δ_i of a random variable δ, and where the random sample $(\delta_1,\dots,\delta_N)$ is supposed to be independent and identically distributed. In a real application, the variable δ describes uncertainty, and δ_i are observations of the variable δ that come from previous experience. The quantity being minimized in (1) is the empirical estimate of a measure of risk, known in financial risk management as *Conditional Value-at-Risk* (CVaR) or *Expected Shortfall* (ES). To make the meaning of (1) concrete, we introduce at this early stage an example that will be resumed later with more explanation and numerical results.

Example 1.1 (Portfolio optimization)**.** Suppose that n_a assets $A^{[1]},\dots,A^{[n_a]}$ are available for trading. On period i, the asset $A^{[j]}$ may gain or lose value in the market, and the ratio $\delta_i^{[j]} = (P_i^{[j]} - P_{i-1}^{[j]})/P_{i-1}^{[j]}$, where $P_i^{[j]}$ is the close price of asset $A^{[j]}$ on period i, is called the *rate of return* of asset $A^{[j]}$ on period i. To cope with uncertainty, investors *diversify* among assets; thus, if an investor has 1$ to invest, s/he will invest fractions $x^{[1]},\dots,x^{[n_a]}$ of her/his dollar on $A^{[1]},\dots,A^{[n_a]}$ (we assume that $x^{[j]} \ge 0$ for all j, and $\sum_{j=1}^{n_a} x^{[j]} = 1$). The vector $x := (x^{[1]},\dots,x^{[n_a]})$ is called a *portfolio*. Letting $\delta_i := (\delta_i^{[1]},\dots,\delta_i^{[n_a]})$ be the vector of the rates of return, the scalar product $\delta_i \cdot x = \sum_{j=1}^{n_a} \delta_i^{[j]} x^{[j]}$ is the rate of return of the portfolio on period i. If $\delta_i \cdot x$ is positive, the investor's capital increases on period i of $\delta_i \cdot x$ \$ for each dollar invested. Hence,

$$L_i(x) := -\delta_i \cdot x$$

quantifies the *portfolio loss* on period i.

Suppose now that the investor has observed a record of N vectors $(\delta_1,\dots,\delta_N)$ on various periods. Then s/he can choose a portfolio x_N^* by minimizing cost (1) where $\mathcal{X} = \{x \in \mathbb{R}^{n_a} : \ x^{[j]} \ge 0 \text{ for all } j, \ \sum_{j=1}^{n_a} x^{[j]} = 1\}$ is the simplex in $\mathbb{R}^{n_a}$. The interpretation is that the investor chooses the portfolio that incurs the lowest average loss over the empirical shortfall cases.

CVaR is a coherent risk measure in the sense of Artzner, Delbaen, Eber, and Heath (1999), which has been introduced and pop-

* Corresponding author.
E-mail addresses: federico.ramponi@unibs.it (F.A. Ramponi), marco.campi@unibs.it (M.C. Campi).

https://doi.org/10.1016/j.ejor.2017.11.022
0377-2217/

ularized by Rockafellar and Uryasev (2000) and Rockafellar and Uryasev (2002) in their papers.[1] ES (see e.g. Christoffersen, 2012 or Fabozzi, Kolm, Pachamanova, & Focardi, 2007) is defined similarly to CVaR, and the difference between CVaR and ES arises only when the distribution of $L(x,\cdot)$ has point masses (that is, there are single values that have non-zero probability to occur). Moreover, in finance, often such a difference is not even considered and a definition of ES completely equivalent to CVaR is used, see e.g. Acerbi and Tasche (2002) and McNeil, Frey, and Embrechts (2015, Chapter 2). In this paper we deal with distributions *without* point masses, and use the definition of ES (or CVaR) that is given in formula (4) below. Terminology and technicalities aside, the concept of expected shortfall is gaining popularity in fields well outside the realm of financial analysis. For example, ES as a measure of risk has recently seen applications to breast cancer therapy (Chan, Mahmoudzadeh, & Purdie, 2014), scheduling (Quan, He, & He, 2014; Sarin, Sherali, & Liao, 2014), and machine learning (Takeda, 2009; Takeda & Kanamori, 2009, 2014; Wang, Dang, & Wang, 2015).[2]

As said before, Problem (1) is a heuristic towards the minimization of an ES risk measure and in this paper we provide rigorous results that certify the properties of this heuristic. More precisely, we introduce a notion of *shortfall threshold* $\bar{L}_N$ (see Eq. (6)) which is interpreted as the empirical boundary of shortfall cases and consider the event where "a further function $L(\cdot,\delta)$, with δ sampled from P independently of the already seen values $(\delta_1,\dots,\delta_N)$, incurs a cost $L(x_N^*,\delta)$ bigger than $\bar{L}_N$". Such a probability is written as $\mathsf{P}\{\delta : L(x_N^*,\delta) > \bar{L}_N\}$, and is a random variable because it depends on x_N^* and $\bar{L}_N$, which in turn depend on the random sample $\delta_1,\dots,\delta_N$.

We show that a probabilistic certificate of the form

$$\mathsf{P}\{\delta : L(x_N^*,\delta) > \bar{L}_N\} \le \varepsilon \quad \text{with confidence } 1-\beta \tag{2}$$

can be attached to the solution of (1). This result has a *universal* validity, that is, it holds true regardless of the distribution P by which the δ_i's are sampled. Hence, an experimenter unaware of P can still append to the solution of Problem (1) a probabilistic certificate in the form of (2). This paper also shows the usefulness of this result by providing a set of corollaries that have a practical use, as well as application examples with real data.

1.1. Structure of the paper

Relevant definitions are given in Section 2. In Section 3 the main result that the random variable $\mathsf{P}\{\delta : L(x_N^*,\delta) > \bar{L}_N\}$ has a universal distribution is stated and proven, followed by two corollaries regarding the statistics and the long-run behavior of such random variable. Section 4 presents two applications exploring, respectively, the choice of k, and the long-run behavior of a sequence of optimization problems solved in a "sliding window" fashion. In Section 5, the results from Sections 3 and 4 are applied to the optimization of a portfolio that includes shares of 10 companies with high market capitalization traded on the New York Stock Exchange and the NASDAQ. The paper ends with some conclusions and acknowledgments.

2. Formal definitions and problem position

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, $(\Delta, \mathcal{F}, \mathsf{P})$ be a probability space, and $L : \mathcal{X} \times \Delta \to \mathbb{R}$ be a function such that

1. For any $x \in \mathcal{X}$, $L(x,\cdot)$ is a random variable on $(\Delta, \mathcal{F}, \mathsf{P})$;
2. For any $\delta \in \Delta$, $L(\cdot,\delta)$ is a convex function on $\mathcal{X}$.

L is interpreted as a cost function whose value depends on an optimization variable x and a variable δ (uncertainty variable) that accounts for all other sources of variation of L besides x. If $(\delta_1,\dots,\delta_N)$ is a sample of independent realizations from $(\Delta, \mathcal{F}, \mathsf{P})$, we shall often use the shorthand notation $L_i := L(\cdot,\delta_i)$, and $L_i(x) := L(x, \delta_i)$, $i = 1,\dots,N$.

For any $x \in \mathcal{X}$, denote by $L_{(i)}(x)$, $i = 1,\dots,N$, the values attained by $L_1(x),\dots,L_N(x)$ taken in descending order:

$$L_{(1)}(x) \ge L_{(2)}(x) \ge \cdots \ge L_{(N)}(x).$$

In statistical terminology, David and Nagaraja (2003), $L_{(N-i+1)}(x)$ is called the ith order statistic of the random sample $L_1(x),\dots,L_N(x)$. Problem (1) can now be restated as follows:

$$\min_{x\in\mathcal{X}} \frac{1}{k}\sum_{i=1}^{k} L_{(i)}(x), \tag{3}$$

where $1 \le k \le N$.

We next introduce a definition of expected shortfall. If L is a random variable modeling a loss, $\alpha \in [0,1]$, and F_L is the cumulative distribution function of L, the *Value at Risk* (VaR) and *Expected Shortfall* (ES) of L are given by:

$$\begin{aligned} \mathrm{VaR}_\alpha(L) &:= \min\{l \in \mathbb{R} : \ F_L(l) \ge \alpha\}, \\ \mathrm{ES}_\alpha(L) &:= \mathsf{E}[L \mid L > \mathrm{VaR}_\alpha(L)]. \end{aligned} \tag{4}$$

$\mathrm{VaR}_\alpha(L)$ is the threshold value at the boundary of the fraction α of highest losses. VaR is currently the most widely adopted risk measure in banking and finance despite some of its shortcomings seem to suggest that it would be better replaced by other measures like ES (refer e.g. to Christoffersen (2012) and Fabozzi et al. (2007) for examples and practical uses, and to Rockafellar and Uryasev (2002) and Hong, Hu, and Liu (2014) for a comparison of the properties of VaR and ES). $\mathrm{ES}_\alpha(L)$ is instead the expected loss suffered when the threshold $\mathrm{VaR}_\alpha(L)$ is exceeded. When the loss L depends on a choice $x \in \mathcal{X}$, i.e., $L = L(x)$, it makes sense to minimize the expected shortfall for a selected value of α:

$$\min_{x\in\mathcal{X}} \mathrm{ES}_\alpha(L(x)). \tag{5}$$

Problem (3) is indeed an empirical version of Problem (5) for $\alpha = 1 - \frac{k}{N}$, based on the N observations $\delta_1,\dots,\delta_N$. Hence, we call Problem (3) the *empirical expected shortfall* problem.

Let x_N^* be the minimizer of (3), assume that it exists and is unique and, assuming also that $N \ge k + d$, define

$$\bar{L}_N := L_{(k+d)}(x_N^*). \tag{6}$$

We call $\bar{L}_N$ the *shortfall threshold*. In typical cases the interpretation of $\bar{L}_N$ is that it separates shortfall empirical functions from functions attaining a lower value at the minimizer. This is easily understood by making reference to a simple case where $d = 1$ and $k = 2$, as shown in Fig. 1(a). The dashed function is $\frac{1}{2}(L_{(1)}(x) + L_{(2)}(x))$. x_N^* minimizes this dashed function, which happens at the intersection of two functions L_i. $\bar{L}_N = L_{(2)}(x_N^*) = L_{(3)}(x_N^*)$ is at the boundary of the values attained by the functions L_i that are averaged to determine the solution. Notice, however, that there are cases where $\bar{L}_N$ takes a value lower than the boundary value. For example, in Fig. 1(b) the solution is determined by two functions only, and $\bar{L}_N$ is obtained by "digging" at x_N^* until the third value $L_{(3)}(x_N^*)$ is reached. This situation may occur when the cost functions are not linear, as in Fig. 1(b), or even when they are linear and the solution x_N^* is obtained at a boundary point of the optimization domain $\mathcal{X}$. The reason why $\bar{L}_N$ is defined to always be the $(k+d)$th largest cost is that the theoretical certificate introduced in this paper holds true rigorously for this choice only.

[1] A recent work of Mafusalov and Uryasev (2016) has generalized the concept of CVaR from that of a risk measure to that of a norm over a space of random variables. In fact the average of the k greatest values among $|l_1|,\dots,|l_N|$ of a vector $(l_1,\dots,l_N) \in \mathbb{R}^N$ is a norm on $\mathbb{R}^N$; for $k = 1$ it reduces to the Chebycheff norm $\|\cdot\|_\infty$.

[2] The minimization problem with empirical distribution in Takeda and Kanamori (2014, Section 3.2) is essentially Problem (1).

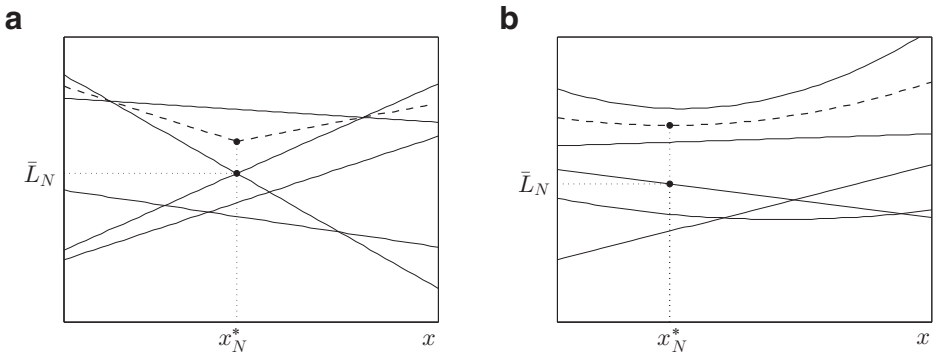

Fig. 1. The shortfall threshold $\bar{L}_N$. The solid lines represent the functions L_i and the dashed line is $\frac{1}{2}(L_{(1)}(x) + L_{(2)}(x))$.

By definition, $k+d$ empirical functions out of N have a value at x^*_N bigger than or equal to $\bar{L}_N$. Moreover, under a non-degeneracy assumption (Assumption 3.2 below), this number is exact in the sense that no other function L_i attains at x^*_N the value $\bar{L}_N$. Hence, the empirical probability of exceeding or meeting the value $\bar{L}_N$ is $(k+d)/N$. In other words, $\bar{L}_N$ is the VaR_α, where $\alpha = 1 - \frac{k+d}{N}$, of a random variable having a discrete distribution function with equal masses placed at $L_1(x^*_N), \ldots, L_N(x^*_N)$. The value $(k+d)/N$ can be seen as an estimate of the probability $\mathsf{P}\{\delta : L(x^*_N, \delta) > \bar{L}_N\}$,[3] which is the true probability with which a next, still unseen, situation δ will incur a cost $L(x^*_N, \delta)$ greater than the shortfall threshold $\bar{L}_N$. This paper establishes results on the true probability that $L(x^*_N, \delta) > \bar{L}_N$, and Theorem 3.1 provides a precise answer that this probability has a Beta distribution irrespective of P.

This result is of great importance in applications, where assuming knowledge of P is often not realistic. This is in contrast with most literature on ES, where it is often assumed that P belongs to a parametric class or that it is however restricted in some specific way: see e.g. Quan et al. (2014) where the inflows have a Gamma distribution, Šutien, Kabašinskas, Strebeika, Kopa, and Reichardt (2014) where the returns possess "α-stable" distributions, Ponomareva, Roman, and Date (2015) where moments are used to generate scenarios for further mean-risk analysis, (Takeda & Kanamori, 2014, Section 3.3) for a case in which the first and second moments are known, Natarajan, Pachamanova, and Sim (2009) and Zymler, Rustem, and Kuhn (2011) dealing with parameter uncertainty by means of robust optimization, or Zhu and Fukushima (2009) for an introduction to the so-called Worst-Case CVaR, whose very definition requires that P belongs to an a-priori known class $\mathcal{P}$.

We feel advisable to spend some further words to highlight some aspects of the mathematical problem dealt with in this paper, that might otherwise be missed. Note first that the problem of evaluating $\mathsf{P}\{\delta : L(x, \delta) > L_{(k+d)}(x)\}$ for a *fixed* x is a standard problem in probability which falls within the frame of order statistics. Specifically, by applying the result in e.g. David and Nagaraja (2003, p. 10), Shao (2003, p. 102), or Gentle (2009, p. 63), one concludes that, if $L(x,\cdot)$ has no point masses, $\mathsf{P}\{\delta : L(x, \delta) > L_{(k+d)}(x)\}$ distributes as a Beta$(k+d, N+1-k-d)$ random variable. On the other hand, in the context of the present paper we deal with the expected shortfall solution x^*_N, which is *not* fixed and depends on the random sample $(\delta_1, \ldots, \delta_N)$. In other words, the x-value at which $\mathsf{P}\{\delta : L(x^*_N, \delta) > L_{(k+d)}(x^*_N)\}$ is evaluated is itself stochastic; as a consequence, the standard theory of order statistics cannot be applied to this context. It is a fact that for a generic choice of a function $x^*_N(\delta_1, \ldots, \delta_N)$, and for a given k, the result that $\mathsf{P}\{\delta : L(x^*_N(\delta_1, \ldots, \delta_N), \delta) > L_{(k+d)}(x^*_N(\delta_1, \ldots, \delta_N))\}$ has a Beta distribution fails to be true, and a simple example showing this fact is provided below.

Example 2.1 (VaR optimization)**.** Let $\mathcal{X} = [0, 4]$, $\delta = (c, \alpha)$, where c is a Bernoullian random variable taking values {0,1} with probability 1/2 each, and α is a uniform random variable over [0,1] independent of c, and let

$$L(x, \delta) = cx + (1-c)\left(2 - \frac{x}{2}\right) + \alpha;$$

see Fig. 2(a). Further, let $N = 3$ and $k = 2$, and

$$x^*_3(\delta_1, \delta_2, \delta_3) = \arg\min_{x \in \mathcal{X}} L_{(2)}(x).$$

In words, x^*_3 is the random point minimizing the level-2 VaR over [0,4]. Only one of two situations may occur: either L_1, L_2, L_3 have all the same slope (see Fig. 2(b)), or two of them have the same slope, while the other has opposite slope (see Fig. 2(c)). A simple reasoning reveals that in both cases either $x^*_3 = 0$ or $x^*_3 = 4$, and in both cases $L_{(3)}(x^*_3) \le 1$. Therefore it holds that $\mathsf{P}\{\delta : L(x^*_3, \delta) > L_{(2+1)}(x^*_3)\} \ge 1/2$, so that in this example the cumulative distribution function of $\mathsf{P}\{\delta : L(x^*_N, \delta) > L_{(k+d)}(x^*_N)\}$ takes the value 0 over [0,1/2), and is not a Beta distribution.

2.1. Comparison with other results in the literature

In the last decade, a new methodology called the *scenario approach* has been introduced where the following min-max program is considered ($\mathcal{X}$ and δ_i have here the same meaning as in (1)):

$$x^*_N = \arg\min_{x \in \mathcal{X}} \max_{i=1,\ldots,N} L(x, \delta_i). \tag{7}$$

Note that (7) is a particular case of (1) obtained for $k = 1$. The minimum attained by x^*_N is denoted by L^*_N.

The scenario approach was firstly introduced in Calafiore and Campi (2006), then studied by others, see e.g. Campi and Garatti (2008), Kanamori and Takeda (2012), Esfahani, Sutter, and Lygeros (2015), and Carè, Garatti, and Campi (2015) for recent developments, and has come in widespread use with many applications to model identification (Campi, Calafiore, & Garatti, 2009) and systems and control engineering (Calafiore & Campi, 2006; Campi, Garatti, & Prandini, 2009). For Problem (7), papers (Calafiore & Campi, 2006; Campi & Garatti, 2008) contain studies of the probability $\mathsf{P}\{\delta : L(x^*_N, \delta) > L^*_N\}$, which has been shown to have a Beta

[3] Under Assumption 3.2, this is the same as $\mathsf{P}\{\delta : L(x^*_N, \delta) \ge \bar{L}_N\}$.

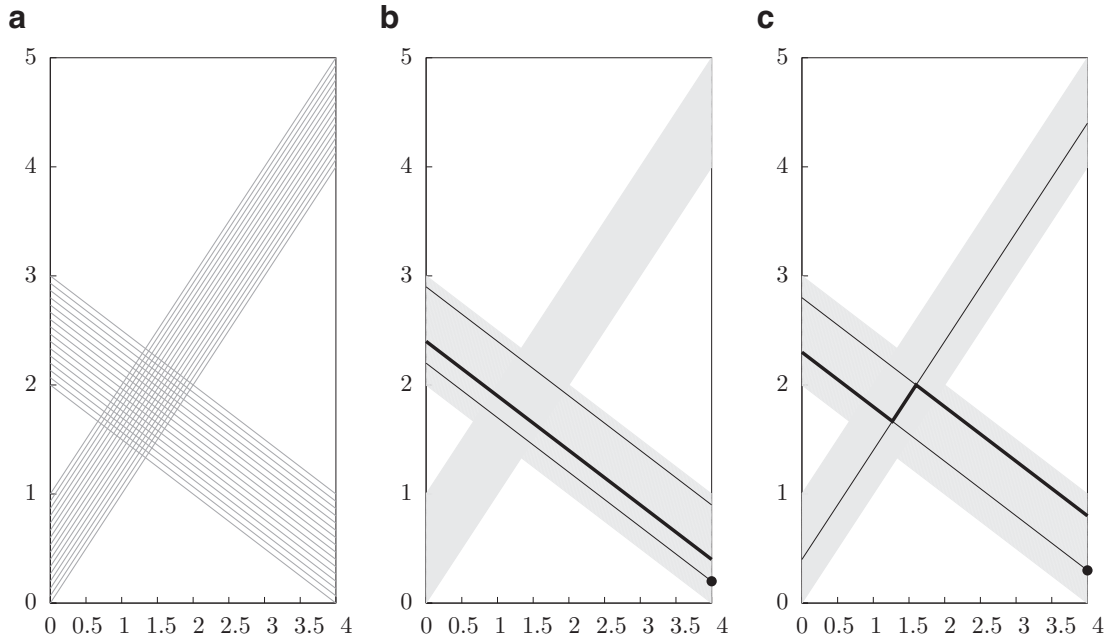

Fig. 2. (a) Functions $L(\cdot,\delta)$; (b) sample of $N = 3$ functions with the same slope (the point $(x_3^*, L_{(2+1)}(x_3^*))$ is marked with •); (c) sample of $N = 3$ functions, two with the same slope and one with opposite slope.

distribution. The results of this paper extend in a non-trivial manner these achievements to expected shortfall optimization.

The worst-case Problem (7) suffers from the inconvenience that its solution is dictated by few ill situations and is therefore quite sensitive to outliers. To cope with this difficulty, in Campi and Garatti (2011) the following problem of the VaR type has been considered:

$$\min_{x\in\mathcal{X}} \{\text{the } k\text{th largest value among } L_1(x), \ldots, L_N(x)\}. \tag{8}$$

The main drawback with (8) is that it is a highly non-convex optimization problem whose exact solution can be hardly found, see e.g. Pagnoncelli, Reich, and Campi (2012) for an application of (8) to portfolio optimization and a discussion on how sub-optimal solutions can be found, and (Natarajan, Pachamanova, & Sim, 2008; Zymler, Kuhn, & Rustem, 2013) for parameter-dependent relaxations of the VaR leading to tractable problems. Moreover, the guarantee provided by Campi and Garatti (2011) for Problem (8) is only a conservative bound.

The expected shortfall studied in this paper avoids the above-mentioned drawbacks that turn up with the VaR approach of Campi and Garatti (2011). In particular, the minimization of the empirical ES is a convex problem. When the cost functions are linear, as is typical in financial mean-risk analysis, it can be solved by means of Linear Programming (see e.g. Mansini, Ogryczak, and Speranza, 2007, Mansini, Ogryczak, and Speranza, 2014, Espinoza and Moreno, 2014; Ponomareva et al., 2015, Bertsimas, Lauprete, and Samarov, 2004, see also Mansini, Ogryczak, and Speranza (2015) for an in-depth survey of LP applied to mean-risk analysis). Moreover, the theoretical results obtained in this paper are tight and then non-conservative.

3. Probabilistic certificates

This section contains the main result that $\mathsf{P}\{\delta : L(x_N^*, \delta) > \tilde{L}_N\}$ has a Beta distribution irrespective of probability P (Theorem 3.1). This result leverages upon Propositions 3.2 and 3.3 where the fundamental properties of ES that underpin the result in Theorem 3.1 are established[4]; in turn, these propositions are based on some preliminary results. The section is closed by two corollaries of Theorem 3.1, that have a practical utility and are used in the next Section 4.

It turns out that studying the properties of Problem (3) requires considering other problems with the same structure as problem (3) which, however, are generated from m functions, instead of N, where m is any integer greater than or equal to k. Correspondingly, consider a sample $(\delta_1, \ldots, \delta_m)$ with $m \geq k$ independent realizations from $(\Delta, \mathcal{F}, \mathsf{P})$ and order $L_i(x)$, $i = 1, \ldots, m$, in descending order: $L_{(1)}(x) \geq L_{(2)}(x) \geq \cdots \geq L_{(m)}(x)$. Then, construct the same problem as in (3), where this time we average the largest among m functions

$$\min_{x\in\mathcal{X}} \frac{1}{k}\sum_{i=1}^{k} L_{(i)}(x). \tag{9}$$

Proposition 3.1. *For any $m \geq k$, (9) is a convex minimization problem.*

Proof. [5] Let $\{i_1, \ldots, i_k\} \subseteq \{1, \ldots, m\}$ be an arbitrary choice of k indices. Since for all $\delta \in \Delta$ the function $L(\cdot,\delta)$ is convex, $\frac{1}{k}\sum_{j=1}^{k} L_{i_j}(\cdot)$ is the average of k convex functions and hence is itself convex. The sum $\frac{1}{k}\sum_{i=1}^{k} L_{(i)}(\cdot)$ in (9) is the point-wise maximum among all the possible $\binom{m}{k}$ choices of k indices from m and is therefore a convex function. □

We make the following assumptions.

Assumption 3.1. For any $m \geq k$, the solution x_m^* to Problem (9) exists and is unique almost surely.

[4] It turns out that similar properties do not hold true for other common risk measures, so that the result of this paper is specific to ES. This is the reason why, for example, for the VaR problem in Section 2 the Beta distribution result is not valid.

[5] Throughout, proofs are provided immediately after each result. However, never in the text reference is made to the proofs, so that the reader can decide to skip reading the proofs without any loss of continuity.

For example, this assumption is verified if $\mathcal{X}$ is compact and $L(\cdot,\delta)$ is almost surely lower semi-continuous and strictly convex over the whole $\mathcal{X}$ (this case is, of course, rather conservative). For $m = N$, Assumption 3.1 says that the solution x_N^* to Problem (3) exists and is unique almost surely.

The next assumption is a non-degeneracy condition.

Assumption 3.2. Consider a sample $(\delta_1, \ldots, \delta_{d+2})$ of independent realizations from $(\Delta, \mathcal{F}, \mathsf{P})$. The event

$$\{\text{there exists an } x \in \mathcal{X} \text{ such that } L_1(x) = L_2(x) = \cdots = L_{d+2}(x)\}$$

has probability zero.

In typical situations, at most $d+1$ functions of a variable $x \in \mathbb{R}^d$ meet at isolated points. Assumption 3.2 rules out the degenerate case where one more function passes through one of the points where the other functions meet. This assumption is satisfied in many applications where the losses are continuous quantities and it is a reasonable modeling simplification even when losses are discrete but fine-grained quantities as is in various financial applications. See also Section 5 for an example.

An immediate consequence of Assumption 3.2 is the following.

Lemma 3.1. *For each $x \in \mathcal{X}$, $L(x,\cdot)$ is a random variable without point masses.*

Proof. Suppose that, for some x and l, $\mathsf{P}\{\delta \in \Delta : L(x,\delta) = l\} = \gamma > 0$, and consider an independent sample $(\delta_1, \ldots, \delta_{d+2})$. By independence, $\mathsf{P}^{d+2}\{L_i(x) = l, i = 1, \ldots, d+2\} = \gamma^{d+2} \neq 0$, which contradicts Assumption 3.2. □

In preparation of the main result, Theorem 3.1, note that Problem (9) can be written in epigraphic form as follows:

$$\begin{aligned} \mathcal{P}_m : \quad & \min_{(x,y)\in\mathcal{X}\times\mathbb{R}} \quad y \\ \text{s.t.} \quad & \frac{1}{k}\sum_{j=1}^{k} L_{i_j}(x) \le y \\ & \text{for any choice of } k \text{ indices } \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, m\}. \end{aligned}$$

Since $\{i_1, \ldots, i_k\}$ is any subset of $\{1, \ldots, m\}$ with cardinality k, $\mathcal{P}_m$ has $\binom{m}{k}$ constraints. The average to be minimized in (9) is the point-wise maximum between all the possible averages of k cost functions, hence the solution of the reformulation $\mathcal{P}_m$ is (x_m^*, y_m^*), where x_m^* is the minimizer of (9) and y_m^* is the optimal value of (9).

Suppose now that $m \ge k+d$ (recall that d is the dimension of the optimization variable x) and let

$$\bar{L}_m := L_{(k+d)}(x_m^*).$$

Hence, $\bar{L}_m$ is the $(k+d)$th value in the sequence

$$L_{(1)}(x_m^*), L_{(2)}(x_m^*), \ldots, L_{(m)}(x_m^*).$$

In principle, more than $k+d$ values in this sequence could be greater than or equal to $\bar{L}_m$. The following result says that this does not happen when Assumptions 3.1 and 3.2 are satisfied.

Proposition 3.2. *Suppose that Assumptions 3.1 and 3.2 hold. Then, almost surely, among the cost functions $L_1, \ldots, L_m$ exactly $k+d$ cost functions attain values $\ge \bar{L}_m$ at x_m^*.*

Proof. Consider a probabilistic outcome where no more than $d+1$ functions have value $\bar{L}_m$ at x_m^* which, by Assumption 3.2, happens with probability 1, and suppose for the sake of contradiction that more than $k+d$ values of $L_i(x_m^*)$ are $\ge \bar{L}_m$. This bears two consequences:

(I) the value $\bar{L}_m$ is taken up by at least two functions at x_m^*;
(II) at least k functions are strictly bigger than $\bar{L}_m$ at x_m^*.

Fact (I) is true because if $\bar{L}_m = L_{(k+d)}(x_m^*)$ is taken up by only one function, then $L_{(k+d+1)}(x_m^*) < L_{(k+d)}(x_m^*) = \bar{L}_m$ and therefore only $k+d$ values of $L_i(x_m^*)$ are greater than or equal to $\bar{L}_m$. Fact (II) is a consequence of the fact that at least $k+d+1$ functions attain values greater than or equal to $\bar{L}_m$ while at most $d+1$ have value $\bar{L}_m$. Note now that y_m^* minimizes the largest mean of k functions so that, if a function L_i attains at x_m^* a value strictly less than $L_{(k)}(x_m^*)$, then any constraint of $\mathcal{P}_m$ containing L_i is not active for problem $\mathcal{P}_m$, and it can be removed from $\mathcal{P}_m$ without changing the solution (x_m^*, y_m^*).[6] Hence, because of (II), we conclude that all functions that at x_m^* have a value $\bar{L}_m$ do not concur in forming the solution. Next, consider problem $\bar{\mathcal{P}}_m$ which is the same as $\mathcal{P}_m$ with the only difference that all the indices i corresponding to the functions for which $L_i(x_m^*) = \bar{L}_m$ have been canceled from $\{1, \ldots, m\}$. The solution of $\bar{\mathcal{P}}_m$ is still x_m^*. Consider now any problem $\bar{\mathcal{P}}$ obtained by leaving out the functions that have as indices any deterministic subset of $\{1, \ldots, m\}$. To fix ideas, say that $\bar{\mathcal{P}}$ contains the indices $\{1, \ldots, p\}$ with $p < m$. Let $\bar{x}$ be the solution of $\bar{\mathcal{P}}$. Due to Lemma 3.1, the functions with indices $p+1, \ldots, m$ do not have point masses corresponding to $x = \bar{x}$. Hence, the event where $L_h(\bar{x}) = L_j(\bar{x})$ for two $h, j \in \{p+1, \ldots, m\}$ has zero probability. Summing over all the problems of the form $\bar{\mathcal{P}}$ (which are finite in number), one sees that zero is the probability of the event $\mathcal{E}$ where there exists a subset of indices from $\{1, \ldots, m\}$ such that, after computing the solution of the corresponding problem $\bar{\mathcal{P}}$, two of the functions that are left out assume the same value corresponding to the solution of $\bar{\mathcal{P}}$. Since for each probabilistic outcome $\bar{\mathcal{P}}_m$ is one of the $\bar{\mathcal{P}}$'s, the event where "the value $\bar{L}_m$ is taken up by at least two functions at x_m^*" (refer to (I)) is a subset of the event $\mathcal{E}$ and has therefore probability zero. Hence, the assumption made for the sake of contradiction holds with probability zero, and this concludes the proof of the proposition. □

Proposition 3.2 gives a rule that almost surely selects $k+d$ indices from $\{1, \ldots, m\}$, namely the indices corresponding to the functions that attain value $\ge \bar{L}_m$. We use the symbol σ_m to explicitly denote this selection function:

$$\begin{aligned} &\sigma_m(\delta_1, \ldots, \delta_m) = \{i_1, \ldots, i_{k+d}\}, \\ &\text{where } i_1 < i_2 < \ldots < i_{k+d}. \end{aligned} \tag{10}$$

The following proposition links $\bar{L}_m$ to the selection operated by σ_m.

Proposition 3.3. *Suppose that a further sample δ_{m+1}, independent of $\delta_1, \ldots, \delta_m$, is sampled from $(\Delta, \mathcal{F}, \mathsf{P})$, and that L_{m+1} is added to the problem so obtaining*

$$\begin{aligned} \mathcal{P}_{m+1} : \quad & \min_{(x,y)\in\mathcal{X}\times\mathbb{R}} \quad y \\ \text{s.t.} \quad & \frac{1}{k}\sum_{j=1}^{k} L_{i_j}(x) \le y \\ & \text{for any choice of } k \text{ indices } \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, m+1\}. \end{aligned}$$

Let (x_{m+1}^*, y_{m+1}^*) be the solution of $\mathcal{P}_{m+1}$. Then, almost surely, the corresponding selection of indices is such that

$$\sigma_{m+1}(\delta_1, \ldots, \delta_{m+1}) \neq \sigma_m(\delta_1, \ldots, \delta_m)$$

if and only if $L_{m+1}(x_m^*) > \bar{L}_m$, and in this case the last index of $\sigma_{m+1}(\delta_1, \ldots, \delta_{m+1})$ is $m+1$.

[6] The fact that if a constraint is not active it can be removed from the problem without changing the solution is because the problem is convex (Proposition 3.1). The results in this paper do not hold, in general, for non-convex L_i's.

Proof. Three situations may happen:

1. $L_{m+1}(x^*_m) < \bar{L}_m$: in this case we have $x^*_{m+1} = x^*_m$, the values $L_{(1)}(x^*_m), \ldots, L_{(k+d)}(x^*_m)$ are the same as those of $\mathcal{P}_m$ by definition of $\bar{L}_m$, and, consequently, $\sigma_{m+1}(\delta_1, \ldots, \delta_{m+1}) = \sigma_m(\delta_1, \ldots, \delta_m)$;
2. $L_{m+1}(x^*_m) = \bar{L}_m$: this event happens with probability zero in view of Lemma 3.1 and is therefore negligible;
3. $L_{m+1}(x^*_m) > \bar{L}_m$: we distinguish two sub-cases:
   3.1 it holds that $x^*_m = x^*_{m+1}$. Then, $L_{m+1}(x^*_{m+1}) = L_{m+1}(x^*_m) > \bar{L}_m$ and $m+1$ is therefore included as the last index of $\sigma_{m+1}(\delta_1, \ldots, \delta_{m+1})$.
   3.2 it holds that $x^*_m \neq x^*_{m+1}$. If in problem $\mathcal{P}_{m+1}$ a function L_i attains at x^*_{m+1} a value strictly less than $L_{(k)}(x^*_{m+1})$, then any constraint of $\mathcal{P}_{m+1}$ containing L_i is not active for problem $\mathcal{P}_{m+1}$, and it can be removed from $\mathcal{P}_{m+1}$ without changing the solution (x^*_{m+1}, y^*_{m+1}) (recall that the problem is convex, Proposition 3.1). Hence $L_{m+1}(x^*_{m+1}) \geq L_{(k)}(x^*_{m+1})$ in problem $\mathcal{P}_{m+1}$ since, otherwise, we would have $x^*_{m+1} = x^*_m$. It follows that $m+1$ is included as the last index of $\sigma_{m+1}(\delta_1, \ldots, \delta_{m+1})$. □

Interestingly, the result in Proposition 3.3 does not hold if instead of $\bar{L}_m$ one considers a threshold which is above the boundary of the shortfall situations because the selection made by the σ function can in this case change even when the new function does not exceed the threshold of interest. As a consequence, the probability of exceeding the threshold is not distributed as a Beta in this case and Theorem 3.1 fails to be true. For example, a easy, but cumbersome, computation shows that in the VaR Example 2.1 if one sets $k = 2$, computes the solution according to (3) and considers the distribution of exceeding the value of the second largest function (instead of the $k+d = 3$-rd largest), this distribution is not a Beta distribution.

We now go back to considering Problem (3) involving N constraints. Introduce the notation

$$\mathrm{PS}_N := \mathrm{P}\left\{\delta \in \Delta : \ L(x^*_N, \delta) > \bar{L}_N\right\},$$

where "PS" means Probability of Shortfall. Since x^*_N depends on $(\delta_1, \ldots, \delta_N)$, so does PS_N, and PS_N is a random variable on $(\Delta^N, \mathcal{F}^N, \mathrm{P}^N)$ taking value in [0,1]. The following fundamental result claims that the cumulative distribution function of PS_N is independent of the problem, i.e., it remains the same, and is therefore known without extra knowledge of the problem, for all problems that satisfy Assumptions 3.1 and 3.2.

Theorem 3.1. *Suppose that Assumptions 3.1 and 3.2 hold. Then,* PS_N *has a* $\mathrm{Beta}(k+d, N+1-k-d)$ *cumulative distribution function:*

$$\mathrm{P}^N\{\mathrm{PS}_N \leq \varepsilon\} = \int_0^\varepsilon \frac{\Gamma(N+1)}{\Gamma(k+d)\Gamma(N+1-k-d)} \ p^{k+d-1}(1-p)^{N-k-d} \ dp,$$

where Γ *is Euler's Gamma function.*

Proof. Consider problem $\mathcal{P}_m$, with $m \geq k+d$. To each sample $(\delta_1, \ldots, \delta_m)$, we associate the indices $\{i_1, \ldots, i_{k+d}\} = \sigma_m(\delta_1, \ldots, \delta_m)$, and group together the samples with the same selection of indices. In this way, Δ^m is partitioned in $\binom{m}{k+d}$ sets $S_1, \ldots, S_{\binom{m}{k+d}}$ up to a zero probability set. Since $\delta_1, \ldots, \delta_m$ are independent and identically distributed, all the sets in the partition have the same probability $1/\binom{m}{k+d}$.

We shall next evaluate the probability of the sets S_i along a different approach, which will allow us to compute quantities of interest by equating the computed probability to $1/\binom{m}{k+d}$. Since the probability of the sets S_i is the same, let us focus on one of them, say the set S_1 corresponding to the indices $1, \ldots, k+d$.

Consider problem $\mathcal{P}_{k+d}$ which contains only the functions $L_1, \ldots, L_{k+d}$, and let F denote the cumulative distribution function of PS_{k+d}. As an intermediate step, we aim to compute F.

For a given selection of $L_1, \ldots, L_{k+d}$, almost surely the next function L_{k+d+1} changes the selection of indices if and only if $L_{k+d+1}(x^*_{k+d}) > \bar{L}_{k+d}$ (refer to Proposition 3.3). Moreover, if the selection of indices is changed after L_{k+d+1} is added, it cannot possibly go back to the initial selection $1, \ldots, k+d$ by adding additional functions $L_{k+d+2}, L_{k+d+3}, \ldots$ for any further change would add to the selection the index of the newly added function. As a consequence, $(\delta_1, \ldots, \delta_m) \notin S_1$. On the other hand, if $L_{k+d+1}(x^*_{k+d}) \leq \bar{L}_{k+d}$ the solution x^*_{k+d+1} does not change. Since $x^*_{k+d+1} = x^*_{k+d}$, by iterating the same argument, one ends up to the conclusion that almost surely $(\delta_1, \ldots, \delta_m) \in S_1$ if and only if $L_{k+d+1}(x^*_{k+d}) \leq \bar{L}_{k+d}, \ldots, L_m(x^*_{k+d}) \leq \bar{L}_{k+d}$. We are now ready to compute the probability of S_1. For a given selection of $\delta_1, \ldots, \delta_{k+d}$, let $p(\delta_1, \ldots, \delta_{k+d})$ be the value of PS_{k+d}. One relation $L_{k+d+i}(x^*_{k+d}) \leq \bar{L}_{k+d}$ holds with probability $1 - p(\delta_1, \ldots, \delta_{k+d})$ and the relations hold simultaneously for all $i = 1, \ldots, m-(k+d)$ with probability $(1 - p(\delta_1, \ldots, \delta_{k+d}))^{m-(k+d)}$ due to independence of $\delta_{k+d+1}, \ldots, \delta_m$. Integrating over all possible values of $p(\delta_1, \ldots, \delta_{k+d})$ gives

$$\begin{aligned}\mathrm{P}^m[S_1] &= \int_{\Delta^{k+d}} (1 - p(\delta_1, \ldots, \delta_{k+d}))^{m-(k+d)} \ d\mathrm{P}(\delta_1, \ldots, \delta_{k+d}) \\ &= \int_0^1 (1-p)^{m-(k+d)} \ dF(p).\end{aligned}$$

Now go back to relation $\mathrm{P}^m[S_1] = \frac{1}{\binom{m}{k+d}}$ to obtain

$$\int_0^1 (1-p)^{m-(k+d)} \ dF(p) = \frac{1}{\binom{m}{k+d}}, \tag{11}$$

which holds for any $m \geq k+d$, and provides all moments of the cumulative distribution function F. Hence, F remains fixed by (11) (see e.g. Shiryaev (1996, ch. 2, sec. 12.9, Corollary 1)). Since by the properties of Euler's Beta function it holds that

$$\binom{m}{k+d}(k+d)\int_0^1 (1-p)^{m-(k+d)} p^{k+d-1} \ dp = 1,$$

we conclude that $F(p) = p^{k+d}$.

Consider now the problem with N functions. Partition Δ^N as before in $\binom{N}{k+d}$ sets $S_1, \ldots, S_{\binom{N}{k+d}}$ up to a zero probability set. It holds that

$$\begin{aligned}&\mathrm{P}^N\{\mathrm{PS}_N \leq \varepsilon\} \\ &= \mathrm{P}^N\left[\{\mathrm{PS}_N \leq \varepsilon\} \cap \bigcup_i S_i\right] \\ &= \sum_i \mathrm{P}^N[\{\mathrm{PS}_N \leq \varepsilon\} \cap S_i] \\ &= \binom{N}{k+d}\mathrm{P}^N[\{\mathrm{PS}_N \leq \varepsilon\} \cap S_1] \\ &= \binom{N}{k+d}\int_{\Delta^N} \mathbb{1}_{\{\mathrm{PS}_N \leq \varepsilon\}}(\delta_1, \ldots, \delta_N)\mathbb{1}_{S_1}(\delta_1, \ldots, \delta_N) d\mathrm{P}(\delta_1, \ldots, \delta_N).\end{aligned}$$

Since over S_1 the solution with all N functions coincides with the solution with only the first $k+d$ functions, the last expression

can further be written as

$$\begin{aligned}
&\binom{N}{k+d}\int_{\Delta^{k+d}} \mathbb{1}_{\{p(\delta_1,\dots,\delta_{k+d})\le\varepsilon\}}(\delta_1,\dots,\delta_{k+d})\\
&\quad\times(1-p(\delta_1,\dots,\delta_{k+d}))^{N-(k+d)}\,d\mathsf{P}(\delta_1,\dots,\delta_{k+d})\\
&=\binom{N}{k+d}\int_0^{\varepsilon}(1-p)^{N-(k+d)}\,dF(p)\\
&=\int_0^{\varepsilon}\frac{N!(k+d)}{(k+d)!(N-k-d)!}\;p^{k+d-1}\;(1-p)^{N-(k+d)}\,dp\\
&=\int_0^{\varepsilon}\frac{\Gamma(N+1)}{\Gamma(k+d)\Gamma(N+1-k-d)}\;p^{k+d-1}\;(1-p)^{N-(k+d)}\,dp.
\end{aligned}$$

This concludes the proof. □

The following corollary gives the probability that, after finding x_N^*, one more cost function associated with δ_{N+1} attains at x_N^* a value that exceeds $\bar{L}_N$.

Corollary 3.1. *Under the hypotheses of Theorem 3.1,*

$$\mathsf{P}^{N+1}\left\{L_{N+1}(x_N^*)>\bar{L}_N\right\}=\mathsf{E}[\mathrm{PS}_N]=\frac{k+d}{N+1}.$$

Proof. In the following computation, the dependence of x_N^* on $(\delta_1,\dots,\delta_N)$ is indicated explicitly:

$$\begin{aligned}
&\mathsf{P}^{N+1}\left\{L_{N+1}(x_N^*(\delta_1,\dots,\delta_N))>\bar{L}_N\right\}\\
&\quad=\int_{\Delta^N}\mathsf{P}\left\{\delta_{N+1}\in\Delta:\ L(x_N^*(\delta_1,\dots,\delta_N),\delta_{N+1})>\bar{L}_N\right\}\\
&\qquad\times d\mathsf{P}^N(\delta_1,\dots,\delta_N)\\
&\quad=\int_{\Delta^N}\mathrm{PS}_N\;d\mathsf{P}^N(\delta_1,\dots,\delta_N)\\
&\quad=\mathsf{E}[\mathrm{PS}_N]\\
&\quad=\frac{k+d}{N+1},
\end{aligned}$$

where the last equality follows from the fact that the expected value of a random variable with distribution $\mathrm{Beta}(k+d,N+1-k-d)$ is $\frac{k+d}{N+1}$. □

Corollary 3.1 states that PS_N is a random variable whose mean is $\frac{k+d}{N+1}$. The following corollary claims that the stochastic variability of PS_N vanishes as N grows unbounded and k is proportional to N.

Corollary 3.2. *Under the hypotheses of Theorem 3.1, if $N\to\infty$ and k grows with N so that $\lim_{N\to\infty}\frac{k}{N}=\varepsilon$, then $\mathrm{PS}_N\to\varepsilon$ in the mean-square sense.*

Proof. Since PS_N has a $\mathrm{Beta}(k+d,N+1-k-d)$ distribution, its mean and variance are, respectively, $\frac{k+d}{N+1}$ and $\frac{k+d}{N+1}\cdot\frac{N+1-k-d}{(N+1)(N+2)}$. Then,

$$\begin{aligned}
&\lim_{N\to\infty}\mathsf{E}\left[(\mathrm{PS}_N-\varepsilon)^2\right]\\
&\quad=\lim_{N\to\infty}\mathsf{E}\left[(\mathrm{PS}_N-\mathsf{E}[\mathrm{PS}_N]+\mathsf{E}[\mathrm{PS}_N]-\varepsilon)^2\right]\\
&\quad=\lim_{N\to\infty}\left(\mathrm{Var}[\mathrm{PS}_N]+2\mathsf{E}[(\mathrm{PS}_N-\mathsf{E}[\mathrm{PS}_N])(\mathsf{E}[\mathrm{PS}_N]-\varepsilon)]\right.\\
&\qquad\left.+(\mathsf{E}[\mathrm{PS}_N]-\varepsilon)^2\right)\\
&\quad=0.
\end{aligned}$$

□

4. Practical uses of the results

In this section, by leveraging on the results in Section 3, we derive some facts that are directly applicable to real problems. We start in Section 4.1 with an explanation of how a suitable trade-off between risk and shortfall threshold can be attained, and proceed in Section 4.2 with an analysis of the performance which is obtained in the long run when the Expected Shortfall scheme is applied repeatedly. The next Section 5 provides experimental results of these two setups with real data.

4.1. Risk/threshold tradeoff

For any *small* $\beta\in(0,1)$, one can compute $p_\beta\in(0,1)$ such that

$$\mathsf{P}^N\left\{\mathrm{PS}_N>p_\beta\right\}=\int_{p_\beta}^1 f(p)\,dp=\beta, \tag{12}$$

where $f(p)$ is the density of a $\mathrm{Beta}(k+d,N+1-k-d)$ random variable. If N is big enough, $f(p)$ is concentrated around its mean $\frac{k+d}{N+1}$ (see Corollaries 3.1 and 3.2), with a thin tail. Therefore, even for very small values of β (say, 10^{-6}), p_β is close enough to the mean. This situation is depicted in Fig. 3. We interpret $1-\beta$ as the confidence that PS_N takes a value in the interval $[0,\,p_\beta]$. More explicitly, the risk that a future cost function L_{N+1} exceeds the level $\bar{L}_N$ is guaranteed to be less than p_β with confidence $1-\beta$. If β is chosen to be very small, $\mathrm{PS}_N\le p_\beta$ holds true with "practical certainty".

Grounded on the above reasoning, to obtain a suitable risk/threshold tradeoff one can use the following procedure: given N cost functions, one solves Problem (3) for different values of k, say $k_1<\dots<k_i<\dots<k_r$. Correspondingly, s/he obtains r different solutions $x_N^{*(1)},\dots,x_N^{*(r)}$, and r different shortfall thresholds $\bar{L}_N^{(1)},\dots,\bar{L}_N^{(r)}$. Then, s/he computes the quantiles $p_\beta^{(i)}$ defined as in Eq. (12). It holds that

$$\begin{aligned}
&\mathsf{P}^N\left\{\mathrm{PS}_N^{(1)}>p_\beta^{(1)}\ \text{ or }\ \dots\ \text{ or }\ \mathrm{PS}_N^{(r)}>p_\beta^{(r)}\right\}\\
&\quad\le\sum_{i=1}^r\mathsf{P}^N\left\{\mathrm{PS}_N^{(i)}>p_\beta^{(i)}\right\}=\sum_{i=1}^r\beta=r\beta.
\end{aligned}$$

Let $\beta':=r\beta$. As we have done above, we interpret $1-\beta'$ as a confidence which, for big N, is sufficiently close to 1 to attain "practical certainty". The shortfall thresholds $\bar{L}_N^{(i)}$, $i=1,\dots,r$, can now be plotted versus the corresponding quantiles $p_\beta^{(i)}$, $i=1,\dots,r$, and one can choose a trade-off solution with her/his rule of preference. In view of the above discussion, the risk that a future cost function L_{N+1} exceeds at *any* of the solutions $x_N^{*(i)}$ the corresponding shortfall threshold $\bar{L}_N^{(i)}$ is less than $p_\beta^{(i)}$ with confidence *at least* $1-\beta'$ ("practical certainty"). In other words, *the whole plot* of $(\bar{L}_N^{(i)},p_\beta^{(i)})$, $i=1,\dots,r$, is guaranteed with "practical certainty", and hence the specific choice $\bar{\imath}$ made is guaranteed that $\mathsf{P}\{L_{N+1}(x_N^{*(\bar{\imath})})>\bar{L}_N^{(\bar{\imath})}\}\le p_\beta^{(\bar{\imath})}$ with "practical certainty". See Section 5 for an application of this setup to a problem in finance.

4.2. Sliding windows

Consider now an application in which a sequence of i.i.d. cost functions $L_1,\dots,L_i,\dots$ is observed progressively in time, and one solves, one after another, a sequence of problems $\{\mathcal{P}_{N,j}\}_{j=0}^{\infty}$ of the form (3), where each problem involves only the N subsequent cost functions $L_{j+1},\dots,L_{j+N}$. The following result ensures that, as time goes on, the proportion of problems in which the next cost function L_{j+N+1} exceeds the "current" shortfall threshold $\bar{L}_{N,j}$ approaches $\frac{k+d}{N+1}$.

Theorem 4.1. *Let $x_{N,j}^*$ be the solution of Problem (3) where, instead of L_i, $i=1,\dots,N$, one uses the cost functions L_i, $i=j+1,\dots,j+N$,*

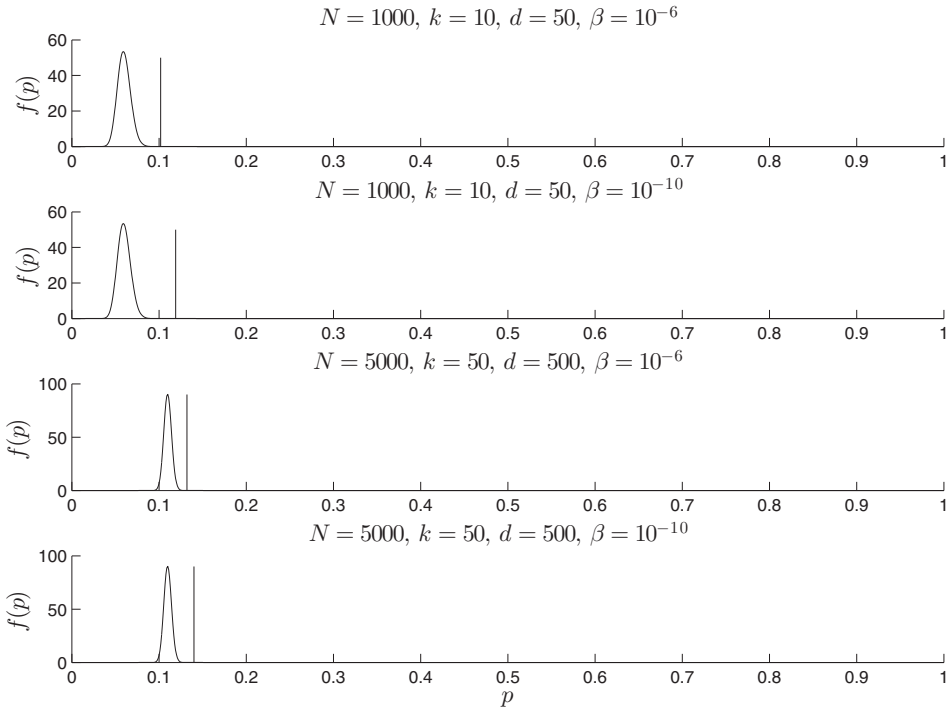

Fig. 3. Quantiles of the Beta distribution. The vertical segment marks the quantile level p_β.

and let $\bar{L}_{N,j}$ be the corresponding shortfall threshold. Then, almost surely,

$$\lim_{T\to\infty} \frac{1}{T}\sum_{j=0}^{T-1} \mathbb{1}_{\{L_{j+N+1}(x^*_{N,j})>\bar{L}_{N,j}\}} = \frac{k+d}{N+1}.$$

Proof. For any T, let τ be the greatest integer such that $\tau(N+1) \le T$, and denote the event $\{L_{j+N+1}(x^*_{N,j}) > \bar{L}_{N,j}\}$ as $\mathcal{E}_j$. Then,

$$\begin{aligned}
&\frac{1}{T}\sum_{j=0}^{T-1} \mathbb{1}_{\{L_{j+N+1}(x^*_{N,j})>\bar{L}_{N,j}\}} = \frac{1}{T}\sum_{j=0}^{T-1} \mathbb{1}_{\mathcal{E}_j} \\
&= \frac{\tau}{T}\left(\frac{1}{\tau}\sum_{i=0}^{\tau-1} \mathbb{1}_{\mathcal{E}_{i(N+1)+0}} + \frac{1}{\tau}\sum_{i=0}^{\tau-1} \mathbb{1}_{\mathcal{E}_{i(N+1)+1}} + \cdots + \frac{1}{\tau}\sum_{i=0}^{\tau-1} \mathbb{1}_{\mathcal{E}_{i(N+1)+N}}\right) \\
&\quad + \frac{1}{T}\sum_{j=\tau(N+1)}^{T-1} \mathbb{1}_{\mathcal{E}_j}.
\end{aligned}$$

Now, each of the $N+1$ terms within parentheses is an average of τ independent and identically distributed random variables. Therefore, since $T\to\infty$ implies $\tau\to\infty$, by the strong law of large numbers almost surely each term tends to $\mathsf{E}[\mathcal{E}_j]$, which equals $\frac{k+d}{N+1}$ (see Corollary 3.1). Moreover, $\frac{1}{T}\sum_{j=\tau(N+1)}^{T-1} \mathbb{1}_{\mathcal{E}_j} \to 0$ and $\frac{\tau}{T} \to \frac{1}{N+1}$. Summing up,

$$\begin{aligned}
&\lim_{T\to\infty} \frac{1}{T}\sum_{j=0}^{T-1} \mathbb{1}_{\{L_{j+N+1}(x^*_{N,j})>\bar{L}_{N,j}\}} \\
&\quad = \frac{1}{N+1}\left(\frac{k+d}{N+1} + \cdots + \frac{k+d}{N+1}\right) + 0 = \frac{k+d}{N+1} \quad \text{almost surely.}
\end{aligned}$$

□

The sequential scheme described above can be refined so that a player can decide whether or not to enter a game depending on observations collected up to the current instant of time $j+N$. Precise probabilistic guarantees hold also in this case. To be specific, suppose that the value of $\bar{L}_{N,j}$ is inspected at each time and, if this value is below a guard level $\bar{G}$, the player enters the game (for example s/he invests, see Section 5 for an application). A posteriori, the game is "won" if it happens that $L_{j+N+1}(x^*_{N,j}) \le \bar{G}$, and it is lost if $L_{j+N+1}(x^*_{N,j}) > \bar{G}$. More generally, the guard level $\bar{G}$ can depend on the observations collected up to time $j+N$ and it can also be time-varying; hence, it will be denoted by $\bar{G}_j$ in what follows. The following corollary provides a guarantee on the long-run average proportion of lost games.

Corollary 4.1. *Almost surely,*

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{j=0}^{T-1} \mathbb{1}_{\{\bar{L}_{N,j}\le\bar{G}_j \text{ and } L_{j+N+1}(x^*_{N,j})>\bar{G}_j\}} \le \frac{k+d}{N+1}. \tag{13}$$

Proof. Note that

$$\{\bar{L}_{N,j} \le \bar{G}_j\} \cap \{L_{j+N+1}(x^*_{N,j}) > \bar{G}_j\} \subseteq \{L_{j+N+1}(x^*_{N,j}) > \bar{L}_{N,j}\},$$

so that

$$\mathbb{1}_{\{\bar{L}_{N,j}\le\bar{G}_j \text{ and } L_{j+N+1}(x^*_{N,j})>\bar{G}_j\}} \le \mathbb{1}_{\{L_{j+N+1}(x^*_{N,j})>\bar{L}_{N,j}\}}.$$

The result now follows by an application of Theorem 4.1. □

Note that Eq. (13) bounds the proportion of times in which the player enters the game and s/he incurs a loss $L_{j+N+1}(x^*_{N,j})$ greater than $\bar{G}_j$. Eq. (13), of course, does not bound the ratio of times in which s/he incurs a loss greater than to $\bar{G}_j$ over the number of times in which s/he plays. If $\bar{G}_j$ is pushed down to low values, for example, the latter ratio can be arbitrarily close to 1.

5. Application example: portfolio optimization

In this section, we consider again the setup of Example 1.1. Theorem 3.1 gives in this context the following result.

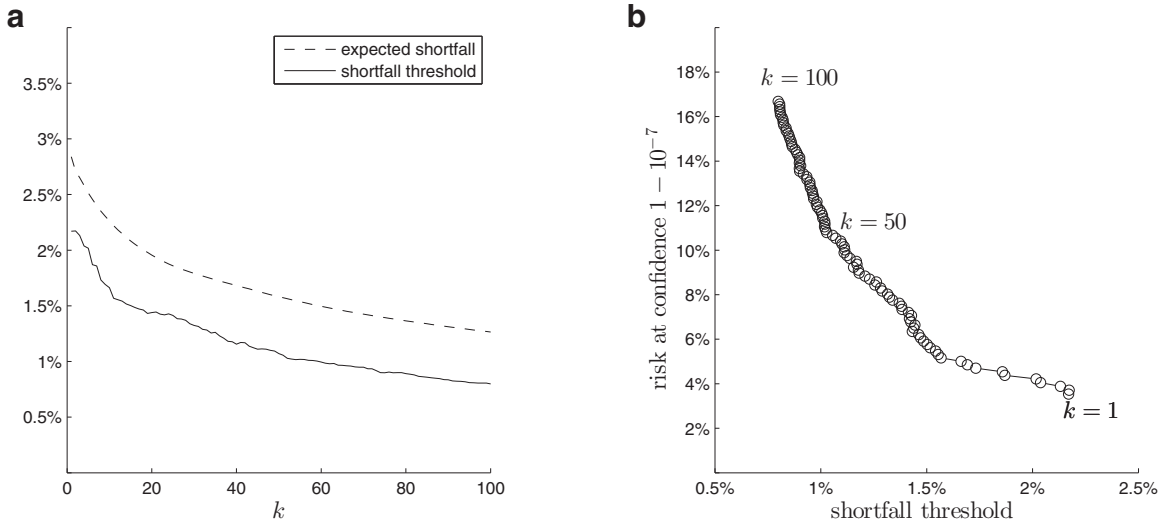

Fig. 4. Risk/threshold tradeoff for a portfolio of $n_a = 10$ assets optimized over $N = 1000$ periods. (a) Plot of the expected shortfall and of the shortfall threshold for the portfolio given by (3) when k ranges from 1 to 100. (b) Shortfall threshold versus risk at confidence $1 - 10^{-7}$.

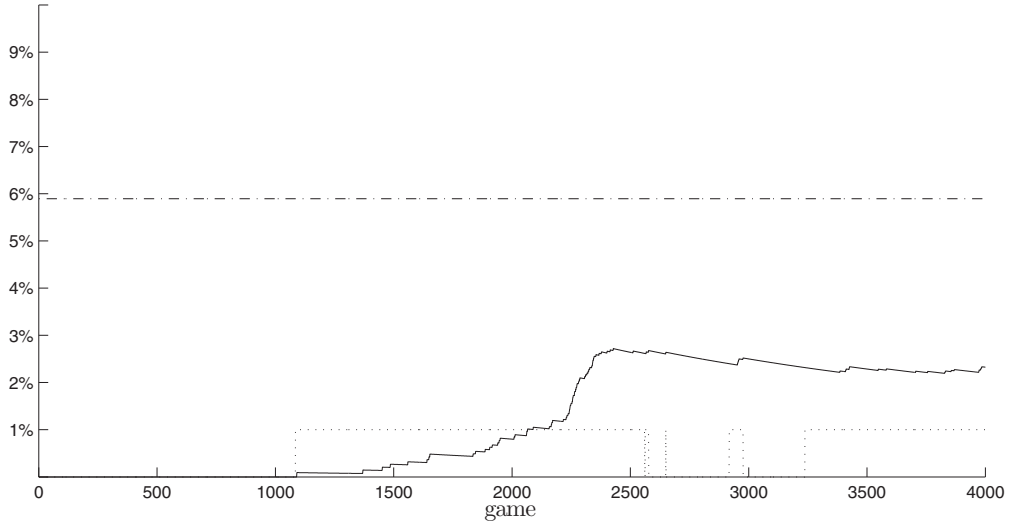

Fig. 5. Sliding window. Dotted line $(\cdot\cdot) = 1$ when game is entered; solid line $(-) =$ average number of lost games; dashed-dotted line $(-\cdot) =$ upper bound in (14) obtained from Corollary 4.1.

Proposition 5.1. *Suppose that $\delta_1, \ldots, \delta_N$ are independent and identically distributed random vectors*[7] *sampled from a probability distribution on $\mathbb{R}^{n_a}$ that admits density*[8]*. Define $\bar{L}_N := L_{(k+n_a-1)}(x_N^*)$. Then the probability of the event $\{L_{N+1}(x_N^*) > \bar{L}_N\}$, i.e., the probability that the future portfolio loss $L_{N+1}(x_N^*)$ is greater than $\bar{L}_N$, has a distribution* $\mathrm{Beta}(k + n_a - 1, N + 2 - k - n_a)$.

Proof. We want to apply Theorem 3.1. Note, first, that the simplex $\mathcal{X}$ is a compact subset of an affine subspace of dimension $d = n_a - 1$ of $\mathbb{R}^{n_a}$, and that the cost functions L_i are affine, and hence convex over $\mathcal{X}$. Therefore, a solution x_N^* always exists. Since, moreover, the vectors of the rates of return $\delta_1, \ldots, \delta_N$ admit density, the piecewise linear function $\frac{1}{k}\sum_{i=1}^{k} L_{(i)}$ is non-flat with probability 1 around its minimizer, so that its minimizer x_N^* is almost surely unique. Moreover, again due to the fact that the distribution of the vectors of the rates of return $\delta_1, \ldots, \delta_N$ admits density, $d + 2$ cost functions L_i pass through the same point with probability 0. Hence, Assumptions 3.1 and 3.2 hold. The result now follows by an application of Theorem 3.1. □

Proposition 5.1 is next illustrated by means of real data.

We consider the 1001 close prices from October 10, 2011 to October 1, 2015 of $n_a = 10$ companies in the S&P500 index.[9] From the close prices, we compute $N = 1000$ vectors of rates of return $\delta_1, \ldots, \delta_{1000}$, and select a portfolio according to (3) for various values of k.

Fig. 4(a) shows, for $k = 1, \ldots, 100$, the optimal value of (3), dashed line, and the corresponding shortfall thresholds $\bar{L}_{1000}^{(1)}, \ldots, \bar{L}_{1000}^{(100)}$, solid line. Proposition 5.1 asserts that, for any fixed k, the distribution of the probability of the event $\{L_{1001}(x_{1000}^{*\,(k)}) > \bar{L}_{1000}^{(k)}\}$ is a $\mathrm{Beta}(k + 9, 992 - k)$. Let $\beta = 10^{-7}$.

[7] The independence of rates of return over disjoint periods (e.g. trading days) is a consequence of the Black–Scholes model often assumed as a hypothesis in the literature. See for example (Hull, 2009, Section 13.3). Instead, the assumption that $\delta_1, \ldots, \delta_N$ are identically distributed is realistic when the market can be assumed stationary over the time frame of observation, which is a limiting assumption.

[8] This assumption is a valid approximation in many cases.

[9] These were the 10 companies in the index with the highest market capitalization at the beginning of 2015, namely `AAPL`, `XOM`, `MSFT`, `JNJ`, `WMT`, `WFC`, `GE`, `PG`, `JPM`, `CVX`.

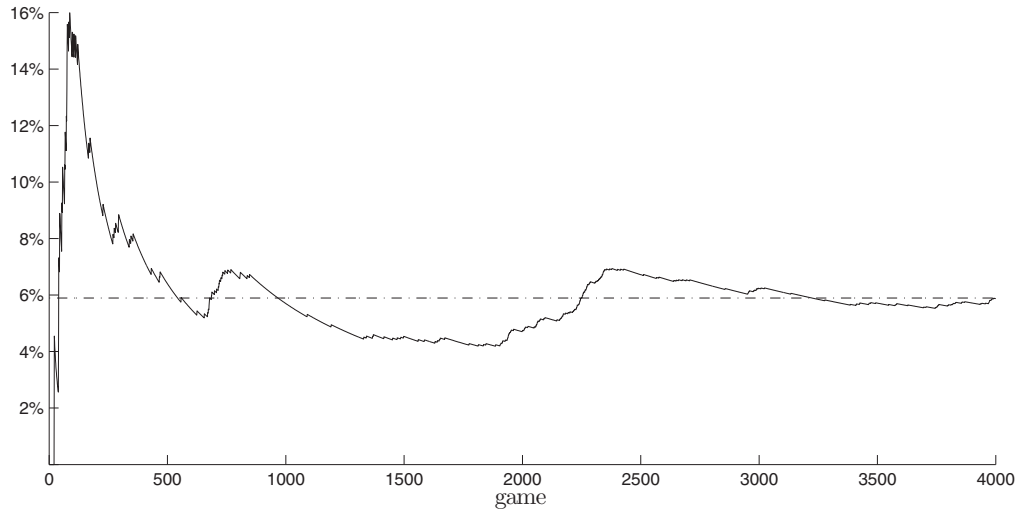

Fig. 6. Sliding window. Solid line (−) = average number of times when $L_{j+N+1}(x^*_{N,j}) > \bar{L}_{N,j}$; dashed-dotted line (−·) = 5.9% obtained from Theorem 4.1.

Fig. 4(b) shows, for each $k = 1, \ldots, 100$, the shortfall threshold $\bar{L}^{(k)}_{1000}$ versus the $(1 - 10^{-7})$-quantile of the corresponding Beta distribution (*risk* at confidence $1 - 10^{-7}$). E.g. for $k = 50$ one reads that the shortfall threshold is $\simeq 1\%$ and the risk is $\simeq 10.5\%$, which is interpreted as a bound on the probability to incur a loss larger than 1%. This plot represents an important source of information that can be used to make a suitable choice of k. According to the theory of Section 4.1, a plot like the one in Fig. 4(b) is guaranteed with a confidence $1 - 100\beta = 1 - 10^{-5}$, so that the risk to incur a loss above the corresponding threshold for the particular choice made is also guaranteed with confidence $1 - 10^{-5}$.

We further consider the sliding window situation described in Section 4.2. This time we consider the 5002 close prices from November 11, 1995 to October 1, 2015 of the same $n_a = 10$ companies as before. We solve, in succession, 4000 optimization problems with $N = 1000$ periods each, and, with $k = 50$, compute $\bar{L}_{N,j}$ for $j = 1, \ldots, 4000$. We take $\bar{G}_j = \bar{G} = 1.5\%$ for any j. If $\bar{L}_{N,j} \leq \bar{G}$, we "enter the game", that is we invest. The game is lost when $L_{j+N+1}(x^*_{N,j}) > \bar{G}$. Corollary 4.1 says that

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{j=0}^{T-1} \mathbb{1}_{\{\text{game is entered at time } j \text{ and game is lost}\}} \leq \frac{k + n_a - 1}{N + 1} \simeq 5.9\%. \tag{14}$$

The solid line in Fig. 5 shows the empirical result when this scheme is applied to the real data.

Before closing this section, we illustrate Theorem 4.1 on the same dataset used for the sliding window example. Fig. 6 gives the average number of times when $L_{j+N+1}(x^*_{N,j}) > \bar{L}_{N,j}$, compared with the value $\frac{k+n_a-1}{N+1} \simeq 5.9\%$. This also corresponds to applying the sliding window scheme with $\bar{G}_j = \bar{L}_{N,j}$, so that one always enters the game.

6. Conclusions

In this paper we have considered the minimization of the empirical Expected Shortfall (ES) computed from a sample of N convex functions defined over a subset of $\mathbb{R}^d$, independently sampled according to a probability P and denoted $L_1(x), \ldots, L_N(x)$. We have shown that $\bar{L}_N =$ the $(k+d)$th order statistic of the sample $L_1(x^*_N), \ldots, L_N(x^*_N)$, where x^*_N is the optimal solution of the ES problem, can be interpreted as a risk threshold. The main result of the paper is that the probability of exceeding $\bar{L}_N$ with a further function L_{N+1} is a random variable with universal distribution $\text{Beta}(k + d, N + 1 - k - d)$ irrespective of P. This result allows one to formulate risk certificates of practical utility. In the last section, these certificates have been applied to an example in portfolio optimization with real financial data.

Acknowledgments

We wish to thank Giorgio Arici, Marco Dalai, Augusto Ferrante, Renata Mansini, and Valerio Volpe for the valuable comments and suggestions they have provided on this manuscript. We extend our thanks to the anonymous reviewers, whose insightful comments have helped us improve the paper.

References

Acerbi, C., & Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance, 26*(7), 1487–1503.
Artzner, P., Delbaen, F., Eber, J. M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance, 9*, 203–228.
Bertsimas, D., Lauprete, G. J., & Samarov, A. (2004). Shortfall as a risk measure: properties, optimization and applications. *Journal of Economic Dynamics & Control, 28*(7), 1353–1382.
Calafiore, G., & Campi, M. C. (2006). The scenario approach to robust control design. *IEEE Transactions on Automatic Control, 51*(5), 742–753.
Campi, M. C., Calafiore, G., & Garatti, S. (2009). Interval predictor models: Identification and reliability. *Automatica, 45*, 382–392.
Campi, M. C., & Garatti, S. (2008). The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization, 19*(3), 1211–1230.
Campi, M. C., & Garatti, S. (2011). A sampling-and-discarding approach to chance–constrained optimization: Feasibility and optimality. *Journal of Optimization Theory and Applications, 148*, 257–280.
Campi, M. C., Garatti, S., & Prandini, M. (2009). The scenario approach for systems and control design. *Annual Reviews in Control, 33*, 149–157.
Carè, A., Garatti, S., & Campi, M. C. (2015). Scenario min-max optimization and the risk of empirical costs. *SIAM Journal on Optimization, 25*(4), 2061–2080.
Chan, T. C. Y., Mahmoudzadeh, H., & Purdie, T. G. (2014). A Robust-CVaR optimization approach with application to breast cancer therapy. *European Journal of Operational Research, 238*(3), 876–885.
Christoffersen, P. F. (2012). *Elements of financial risk management* (2nd ed.). Elsevier.
David, H. A., & Nagaraja, H. N. (2003). *Order statistics* (3rd ed.). Wiley.
Esfahani, P. M., Sutter, T., & Lygeros, J. (2015). Performance bounds for the scenario approach and an extension to a class of non-convex programs. *IEEE Transaction on Automatic Control, 60*(1), 46–58.
Espinoza, D., & Moreno, E. (2014). A primal-dual aggregation algorithm for minimizing conditional-value-at-risk in linear programs. *Computational Optimization and Applications, 59*(3), 617–638.
Fabozzi, F. J., Kolm, P. N., Pachamanova, D., & Focardi, S. M. (2007). *Robust portfolio optimization and management*. Wiley.
Gentle, J. E. (2009). *Computational statistics*. Springer.

Hong, L. J., Hu, Z., & Liu, G. (2014). Monte carlo methods for value-at-risk and conditional value-at-risk: A review. *ACM Transactions on Modeling and Computer Simulation, 24*(4), 22:1–22:37.
Hull, J. (2009). *Options, futures and other derivatives* (8th ed.). Pearson/Prentice Hall.
Kanamori, T., & Takeda, A. (2012). Worst-case violation of sampled convex programs for optimization with uncertainty. *Journal of Optimization Theory and Applications, 152*(1), 171–197.
Mafusalov, A., & Uryasev, S. (2016). CVaR (superquantile) norm: Stochastic case. *European Journal of Operational Research, 249*, 200–208.
Mansini, R., Ogryczak, W., & Speranza, M. G. (2007). Conditional value at risk and related linear programming models for portfolio optimization. *Annals of Operations Research, 152*, 227–256.
Mansini, R., Ogryczak, W., & Speranza, M. G. (2014). Twenty years of linear programming based portfolio optimization. *European Journal of Operational Research, 234*, 518–535.
Mansini, R., Ogryczak, W., & Speranza, M. G. (2015). *Linear and mixed integer programming for portfolio optimization*. Springer International Publishing.
McNeil, A. J., Frey, R., & Embrechts, P. (2015). *Quantitative risk management: Concepts, techniques and tools* (Revised ed.). Princeton University Press.
Natarajan, K., Pachamanova, D., & Sim, M. (2008). Incorporating asymmetric distributional information in robust value-at-risk optimization. *Management Science, 54*(3), 573–585.
Natarajan, K., Pachamanova, D., & Sim, M. (2009). Constructing risk measures from uncertainty sets. *Operations Research, 57*(5), 1129–1141.
Pagnoncelli, B. K., Reich, D., & Campi, M. C. (2012). Risk-return trade-off with the scenario approach in practice: A case study in portfolio selection. *Journal of Optimization Theory and Applications, 155*, 707–722.
Ponomareva, K., Roman, D., & Date, P. (2015). An algorithm for moment-matching scenario generation with application to financial portfolio optimization. *European Journal of Operational Research, 240*, 678–687.
Quan, Y., He, L., & He, M. (2014). Short-term hydropower scheduling with gamma inflows using CVaR and Monte Carlo simulation. In *Proceedings of the seventh international symposium on computational intelligence and design (ISCID)* (pp. 108–111).
Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk, 2*, 21–41.
Rockafellar, R. T., & Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance, 26*, 1443–1471.
Sarin, S. C., Sherali, H. D., & Liao, L. (2014). Minimizing conditional-value-at-risk for stochastic scheduling problems. *Journal of Scheduling, 17*(1), 5–15.
Shao, J. (2003). *Mathematical statistics* (2nd ed.). Springer-Verlag.
Shiryaev, N. (1996). *Probability*. New York: Springer-Verlag.
Takeda, A. (2009). Generalization performance of ν-support vector classifier based on conditional value-at-risk minimization. *Neurocomputing, 72*(10–12), 2351–2358.
Takeda, A., & Kanamori, T. (2009). A robust approach based on conditional value-at-risk measure to statistical learning problems. *European Journal of Operational Research, 198*(1), 287–296.
Takeda, A., & Kanamori, T. (2014). Using financial risk for analyzing generalization performance of machine learning models. *Neural Networks, 57*, 29–38.
Šutien, K., Kabašinskas, A., Strebeika, D., Kopa, M., & Reichardt, R. (2014). Estimation of VAR and CVAR from financial data using simulated alpha-stable random variables. In *Proceedings of the twenty eighth European simulation and modelling conference (ESM)* (pp. 159–163).
Wang, Y., Dang, C., & Wang, S. (2015). Robust novelty detection via worst case CVaR minimization. *IEEE Transaction on Neural Networks and Learning Systems, 26*(9), 2098–2110.
Zhu, S., & Fukushima, M. (2009). Worst-case conditional value-at-risk with application to robust portfolio management. *Operations Research, 57*(5), 1155–1168.
Zymler, S., Kuhn, D., & Rustem, B. (2013). Worst-case value at risk of nonlinear portfolios. *Management Science, 59*(1), 172–188.
Zymler, S., Rustem, B., & Kuhn, D. (2011). Robust portfolio optimization with derivative insurance guarantees. *European Journal of Operational Research, 210*(2), 410–424.