A Theory of the Risk for Empirical CVaR with Application to Portfolio Selection*

ARICI Giorgio · CAMPI Marco C. · CARÈ Algo · DALAI Marco · RAMPONI Federico A.

DOI: 10.1007/s11424-021-1229-3
Received: 29 June 2021
©The Editorial Office of JSSC & Springer-Verlag GmbH Germany 2021

Abstract When decisions are based on empirical observations, a trade-off arises between flexibility of the decision and ability to generalize to new situations. In this paper, we focus on decisions that are obtained by the empirical minimization of the Conditional Value-at-Risk (CVaR) and argue that in CVaR the trade-off between flexibility and generalization can be understood on the ground of theoretical results under very general assumptions on the system that generates the observations. The results have implications on topics related to order and structure selection in various applications where the CVaR risk-measure is used. A study on a portfolio optimization problem with real data demonstrates our results.

Keywords Distribution-free results, empirical CVaR, generalization, order selection, risk, scenario approach.

1 Introduction

Decision processes are often driven by data. In these cases, it is crucial to keep control on the discrepancy between the empirical performance, which is measured on the data set, and the actual performance, which can only be estimated. This task is difficult in general because an optimization process intertwines with the estimation problem, see, e.g., [1].

When the decision-making procedure can be reduced to solving a convex optimization problem where observations act as constraints, strong theoretical guarantees on the statistics of "bad events" (situations in which the decision underperforms) can be established under fairly mild assumptions: The study of these guarantees is the subject of the so-called "scenario approach", see [2–6].

A prominent aspect of the scenario theory is the relation between the "complexity" of the solution and the risk, defined as the probability of bad events. In some contexts, the complexity

ARICI Giorgio · CAMPI Marco C. · CARÈ Algo · DALAI Marco · RAMPONI Federico A.
Department of Information Engineering, University of Brescia, via Branze 38, 25123 *Brescia, Italy.*
Email: g.arici005@unibs.it; marco.campi@unibs.it; algo.care@unibs.it; marco.dalai@unibs.it; federico.ramponi@unibs.it.

*This research was partially supported by Regione Lombardia under Grant MoSoRe E81B19000840007.

◇ *This paper was recommended for publication by Guest Editor ZHANG Ji-Feng.*

is known *a priori*, for example it is set by the number of optimization variables, see, e.g., [7, 8]. In other cases, the complexity of the solution can only be evaluated a posteriori, i.e., after having solved the optimization problem, see [9, 10].

The main objective of this paper is to explore how the deep-seated results of the scenario theory developed in [4, 8, 11] can be used as a tool for performance control in data-driven Conditional Value-at-Risk (CVaR) decision-making. CVaR, see [12–14], is a risk measure born in the context of financial portfolio optimization that has gained attention in various contexts in recent years due to its remarkable properties, foremost the fact of being a *coherent* risk measure, see, e.g., [13, 15, 16]. This in particular implies that CVaR leads to convex optimization problems under mild assumptions. We present a theory that allows the user to keep control on the actual performance, beyond what is empirically observed on the data set. Importantly, this theory plays a key role to select a suitable flexibility of the domain in which the solution is sought. While the results of this paper bear a promise of general applicability, a particular emphasis is put on portfolio optimization. In this setting, we investigate how the number of assets in the portfolio shapes the trade-off between empirical performance and the confidence on the fact that certain significant loss thresholds will not be exceeded.

The paper is structured as follows. In Section 2, we consider data-driven CVaR optimization and study the relation between flexibility and performance; this section builds on previous achievements in [8] and [11]. While portfolio selection is employed as a running example throughout the paper, Section 3 more specifically focuses on this application and offers a study with real historical data. Conclusions are drawn in Section 4.

2 Guaranteed CVaR Minimization

Let us first recall the definition of the Conditional Value-at-Risk measure. Let $\boldsymbol{L}$ be a random variable, representing a loss, defined over a probability space $(\Delta, \mathcal{F}, \mathsf{P})$, and let $F_{\boldsymbol{L}}$ denote the cumulative distribution function of $\boldsymbol{L}$.

For any fixed $\alpha \in (0,1)$, the *Value-at-Risk* (VaR) of $\boldsymbol{L}$ at level α (also known as the α-quantile, or "inverse cumulative distribution") is the quantity

$$\mathrm{VaR}_\alpha(\boldsymbol{L}) = \min\{l : \ F_{\boldsymbol{L}}(l) \geq \alpha\}. \tag{1}$$

Hence, $\mathrm{VaR}_\alpha(\boldsymbol{L})$ is the loss threshold that is exceeded with probability at most $1-\alpha$. The CVaR at level α is defined as

$$\mathrm{CVaR}_\alpha(\boldsymbol{L}) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_a(\boldsymbol{L})\ da. \tag{2}$$

If $\boldsymbol{L}$ happens to be a *continuous* random variable, then $\mathrm{CVaR}_\alpha(\boldsymbol{L})$ is equal to the following quantity:

$$\mathrm{ES}_\alpha(\boldsymbol{L}) = \mathbb{E}\left\{\boldsymbol{L} \,|\, \boldsymbol{L} \geq \mathrm{VaR}_\alpha(\boldsymbol{L})\right\}. \tag{3}$$

In words, $\mathrm{ES}_\alpha(\boldsymbol{L})$ is the expected loss suffered when $\mathrm{VaR}_\alpha(\boldsymbol{L})$ is reached or exceeded, and happens to be the original reason for another name, *Expected Shortfall*, by which the CVaR is also known in the literature. For non-continuous distributions, $\mathrm{CVaR}_\alpha(\boldsymbol{L})$ and $\mathrm{ES}_\alpha(\boldsymbol{L})$ are not equivalent. In fact, if the distribution of $\boldsymbol{L}$ has a concentrated mass precisely at $\mathrm{VaR}_\alpha(\boldsymbol{L})$, then the definition (2) implies a suitable "split" of the concentrated mass, in such a way that exactly a probability $1-\alpha$ is taken into account in averaging, a "split" that definition (3) (or a similar definition with $>$ instead of $\geq$) cannot capture. We also note that, in some recent literature, *Expected Shortfall* and *Conditional Value-at-Risk* are considered synonyms, the definition (2) being the common one for both the names. CVaR was introduced as a measure of risk in the field of financial analysis, [12], but then it attracted interest across a large variety of fields, ranging from machine learning, [17], to medical applications, [18], and the control of extreme events such as river floods, [19]. See [20] for a recent survey.

In several cases, the loss $\boldsymbol{L}$ depends on a decision variable $x \in \mathcal{X}$, and minimizing the CVaR (i.e., opting for the decision x^* that minimizes the average loss in the $(1-\alpha)$-fraction of worst cases) is a suitable criterion for decision-making. This leads to the CVaR minimization problem:

$$x^* = \operatorname*{argmin}_{x \in \mathcal{X}} \ \mathrm{CVaR}_\alpha(\boldsymbol{L}(x)). \tag{4}$$

Clearly, in order to solve the problem (4), the probability distribution $F_{\boldsymbol{L}(x)}(l)$ of $\boldsymbol{L}(x)$ that are obtained as x varies in $\mathcal{X}$ must be available. In most applications, however, assuming this knowledge is not realistic, and decisions are rather made based on a collection of observations, e.g., coming from historical series. This leads to the *Empirical CVaR* problem, which we introduce in the next subsection.

2.1 Empirical CVaR

For any x, $\boldsymbol{L}(x)$ is a random variable over $(\Delta, \mathcal{F}, \mathsf{P})$. A $\delta \in \Delta$ represents uncertainty and, when we want to indicate the value that $\boldsymbol{L}(x)$ takes corresponding to a specific δ, we write $\boldsymbol{L}(x,\delta)$. Throughout, $\boldsymbol{L}(\cdot,\cdot)$ is assumed to be known, so that, if δ is observed, then the value of $\boldsymbol{L}(x,\delta)$ can be computed for all $x \in \mathcal{X}$. This assumption is realistic and natural in several contexts, including the Portfolio Optimization problem that we shall amply consider in this paper. Situations of partial knowledge of $\boldsymbol{L}$ or partial observability of δ are more challenging and fall beyond the scope of this paper.

Portfolio Optimization 1 Consider the problem where $x = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$ represents a portfolio over n assets (i.e., x_i is the percentage of capital invested on the asset i), so that $x_i \in [0,1]$, $i = 1, 2, \cdots, n$, $\sum_{i=1}^n x_i = 1$, and the outcome $\delta = (\delta_1, \delta_2, \cdots, \delta_n)$ is an n-dimensional real vector whose components are the rates of return of the assets in a given day (i.e., the closing price at the current day minus the closing price of the same asset in the preceding day over the closing price in the preceding day). The loss incurred by the portfolio x can be computed as $\boldsymbol{L}(x,\delta) = -x^{\mathrm{T}}\delta$, that is, $\boldsymbol{L}(\cdot,\cdot)$ is a known function. $*$

The data record available to a decision maker is, from now on, identified with a sample of outcomes $(\delta^{(1)}, \delta^{(2)}, \cdots, \delta^{(N)}) \in \Delta^N$. An empirical version of the problem (4) can be formulated as the problem of minimizing the average of the k largest values among $\boldsymbol{L}(x,\delta^{(1)}), \boldsymbol{L}(x,\delta^{(2)}), \cdots,$ $\boldsymbol{L}(x,\delta^{(N)})$, where k has to be suitably chosen, the most natural choice being such that $\alpha =$

$1 - k/N$. The solution to this problem is here denoted by $\boldsymbol{x}_N^*$, i.e.,

$$\boldsymbol{x}_N^* = \operatorname*{argmin}_{x \in \mathcal{X}} \left\{ \text{average of the } k \text{ largest values among } \boldsymbol{L}(x,\delta^{(1)}), \boldsymbol{L}(x,\delta^{(2)}), \cdots, \boldsymbol{L}(x,\delta^{(N)}) \right\}. \quad (5)$$

Note that, differently from the solution of (4), $\boldsymbol{x}_N^*$ is a *random* vector over Δ^N, because it depends on the outcomes $\delta^{(1)}, \delta^{(2)}, \cdots, \delta^{(N)}$.

The Empirical CVaR problem can be formulated more explicitly by defining, for all $x \in \mathcal{X}$, $\boldsymbol{L}_{(1)}(x), \boldsymbol{L}_{(2)}(x), \cdots, \boldsymbol{L}_{(N)}(x)$ as the values $\boldsymbol{L}(x, \delta^{(1)}), \boldsymbol{L}(x, \delta^{(2)}), \cdots, \boldsymbol{L}(x, \delta^{(N)})$ sorted in decreasing order (with repeats). Thus, we have $\boldsymbol{L}_{(1)}(x) \geq \boldsymbol{L}_{(2)}(x) \geq \cdots \geq \boldsymbol{L}_{(N)}(x)$, and the solution to the Empirical CVaR Problem can be written as

$$\boldsymbol{x}_N^* = \operatorname*{argmin}_{x \in \mathcal{X}} \frac{1}{k} \sum_{i=1}^{k} \boldsymbol{L}_{(i)}(x) \ , \quad (6)$$

where $1 \leq k \leq N$.

Portfolio Optimization 2 Note that the portfolio optimization problem with a portfolio of size n can be formulated as a problem with $d = n - 1$ decision variables. In fact $\mathcal{X}$ can be defined as $\mathcal{X} = \{x \in \mathbb{R}^{n-1} : \sum_{i=1}^{n-1} x_i \leq 1, x_i \geq 0, i = 1, 2, \cdots, n-1\}$, and the proportion invested in the n-th asset can be obtained as $x_n = 1 - \sum_{i=1}^{n-1} x_i$. Problem (5) in this case can also be written as a linear problem in epigraphic form:

$$\begin{aligned} &\min\nolimits_{x \in \mathcal{X}, \lambda \in \mathbb{R}} \lambda \\ &\text{subject to:} \quad \lambda \geq -\frac{1}{k} \sum_{j=1}^{k} \left[\sum_{\ell=1}^{n-1} x_\ell \delta_\ell^{(i_j)} + \left(1 - \sum_{\ell=1}^{n-1} x_\ell \right) \delta_n^{(i_j)} \right], \\ &\qquad\qquad \text{for any choice of k indices } \{i_1, i_2, \cdots, i_k\} \subseteq \{1, 2, \cdots, N\}. \end{aligned} \quad (7)$$

*

After the minimizer of (6) is computed (we assume that the minimizer exists and is unique), each value $\boldsymbol{L}_{(i)}(\boldsymbol{x}_N^*)$, $i = 1, 2, \cdots, N$, can be thought of as the empirical Value-at-Risk at level $1 - \frac{i-1}{N}$ of the loss at the computed solution. The question immediately arises as to whether these empirical values reflect the true probability distribution of the loss at the computed solution. In the next section, we address this question and characterize the generalization properties of the empirical solution.

2.2 Statistical Framework

The following assumption is in force throughout the paper.

Assumption 1 (Independence and identical distribution) The observations $\delta^{(1)}$, $\delta^{(2)}$, $\cdots$, $\delta^{(N)}$ are independent and distributed according to the same (unknown) probability law P (i.i.d. observations). In other words, $(\delta^{(1)}, \delta^{(2)}, \cdots, \delta^{(N)})$ is an outcome in the probability space $(\Delta^N, \mathcal{F}^N, \mathsf{P}^N)$, where $\mathcal{F}^N$ is the N-fold product σ-algebra and P^N is the N-fold product measure. *

We briefly discuss the validity of this assumption in the case of portfolio optimization.

Portfolio Optimization 3 The observations $\delta^{(i)}$ for $i = 1, 2, \cdots, N$ represent the return vectors of the assets in the portfolio in the preceding N days, and $\delta_j^{(i)}$ is the rate of return of the j-th asset on the i-th day. The independence of the rates of return over disjoint periods is a typical modeling assumption, see, e.g., [21], Subsection 14.3. On the other hand, assuming that the rates of return are identically distributed is realistic for limited periods, within which the market can be assumed to be stationary. Note also that, on a given day i, various rates of return $\delta_j^{(i)}$, $j = 1, 2, \cdots, n$, can be arbitrarily correlated. $*$

To set the stage, consider first the extreme situation where the optimization domain is restricted to a single value of x, e.g., $\mathcal{X} = \{\overline{x}\}$. In this case, no optimization is really performed, as $\boldsymbol{x}_N^*$ is *a priori* known to be $\boldsymbol{x}_N^* = \overline{x}$. Thus, the values $\boldsymbol{L}_{(1)}(\boldsymbol{x}_N^*), \boldsymbol{L}_{(2)}(\boldsymbol{x}_N^*), \cdots, \boldsymbol{L}_{(N)}(\boldsymbol{x}_N^*)$ are nothing but the ordered values of an i.i.d. sample of real random variables. In this case, the probability that a new outcome δ incurs a loss higher than the largest i-th of the N previously observed values can be easily studied by resorting to the theory of order statistics, [22]. However, when optimization is performed over a nontrivial set $\mathcal{X}$, the framework of order statistics is not useful anymore. In fact, order statistics are ordered values from the real line, while, in our setting, the empirical loss values lie on a random line passing through $\boldsymbol{x}_N^*$, which is selected by solving an optimization problem (see also the discussion in [23], Subsection 1.1). A moment's reflection reveals that the size of the decision space must play a role in determining to what extent the empirical values are representative of the true distribution of the loss. In fact, the more freely the decision variable is allowed to range during the optimization process, the more biased towards small values the empirical loss will be, hence the need of theoretical instruments to control this effect[†].

For example, in portfolio optimization, one can easily observe that, as the size n of the portfolio increases by including more and more assets, the empirical performance improves but, at the same time, it becomes less representative of the future performance. In what follows, we present tools that have their natural ground in the theory of the scenario approach

- to compute upper- (and lower-) bounds to the probability that meaningful thresholds on the loss are exceeded when the solution $\boldsymbol{x}_N^*$ is applied; and
- to drive the user towards a selection of the flexibility to meet a suitable trade-off between empirical performance and loss control.

2.3 Fundamental Results from [8]

In this subsection, we summarize results from [8]; the subsequent Subsections 2.4 and 2.5 of this paper build upon these results. We need some additional assumptions.

Assumption 2 (Convexity) $\boldsymbol{L}(\cdot, \delta)$ is convex on the set $\mathcal{X} \subseteq \mathbb{R}^d$, which is itself assumed convex, for every $\delta \in \Delta$. $*$

[†] This phenomenon is much related to what in machine learning is known as *data overfitting*, and to the *complexity* (*or capacity*) *control* issue, see, e.g., [1].

This assumption is, e.g., satisfied in our running example of Portfolio Optimization, where the loss function $\boldsymbol{L}(x,\delta) = -x^{\mathrm{T}}\delta$ is linear in x. Note, however, that we do not make any assumption about the dependence $\delta \mapsto \boldsymbol{L}(\cdot,\delta)$, which can be arbitrarily complex.

A direct consequence of Assumption 2 is the following proposition.

Proposition 1 *For any k, (6) is a convex minimization problem.* $*$

Proposition 1 is well-known in the CVaR literature, see, e.g., [12]. It easily follows from observing that the function that is minimized in (6) is convex because it is the point-wise maximum among the $\binom{N}{k}$ averages of k functions that can be obtained out of the N functions $\boldsymbol{L}(x,\delta^{(1)}), \boldsymbol{L}(x,\delta^{(2)}), \cdots, \boldsymbol{L}(x,\delta^{(N)})$ (recall that the function $\boldsymbol{L}(\cdot,\delta)$ is convex and any average of convex functions is itself convex).

Assumption 3 **(Existence and uniqueness)** For any k, the solution $\boldsymbol{x}_N^*$ to Problem (6) exists and is unique almost surely. $*$

The next assumption, taken from [8], requires that at most $d+1$ loss functions, as functions of $x \in \mathbb{R}^d$, meet at isolated points. This is normally satisfied when the loss is continuously distributed, and is often a reasonable modeling simplification when losses are discrete but fine-grained quantities. A simple sufficient condition for Assumption 4 to hold in *portfolio optimization* is provided in Appendix 4.

Assumption 4 **(Non-degeneracy)** Suppose that $\delta^{(1)}, \delta^{(2)}, \cdots, \delta^{(d+2)}$ are independent and distributed according to probability P. The event

$$\Big\{\text{there exists an } x \in \mathcal{X} \text{ such that } \boldsymbol{L}(x,\delta^{(1)}) = \boldsymbol{L}(x,\delta^{(2)}) = \cdots = \boldsymbol{L}(x,\delta^{(d+2)})\Big\} \tag{8}$$

has probability zero. $*$

Under Assumptions 2–4, the following Proposition 2 holds, see [8].

Proposition 2 *Let $N \geq k+d$. Almost surely, among the cost functions $\boldsymbol{L}(\cdot,\delta^{(1)})$, $\boldsymbol{L}(\cdot,\delta^{(2)})$, $\cdots$, $\boldsymbol{L}(\cdot,\delta^{(N)})$ exactly $k+d$ of them attain a value greater than or equal to $\boldsymbol{L}_{(k+d)}(\cdot)$ at $\boldsymbol{x}_N^*$.* $*$

Let us consider the $k+d$ indices $i_1, i_2, \cdots, i_{k+d}$ from $\{1, 2, \cdots, N\}$ corresponding to the functions that attain a value at $\boldsymbol{x}_N^*$ greater than or equal to $\boldsymbol{L}_{(k+d)}(\boldsymbol{x}_N^*)$ (Proposition 2 ensures that they are well-defined). It can be shown that the observations $\delta^{(i_1)}, \delta^{(i_2)}, \cdots, \delta^{(i_{d+k})}$ corresponding to these $k+d$ highest-valued losses play the important role of "support observations": If only these $k+d$ observations are kept while the other $N-(k+d)$ are discarded, the solution $\boldsymbol{x}_N^*$ to the Empirical CVaR problem does not change, and the value of $\boldsymbol{L}_{(k+d)}(\boldsymbol{x}_N^*)$ remains the same. On the other hand, removing some of these support observations causes the value $\boldsymbol{L}_{(k+d)}(\boldsymbol{x}_N^*)$ to change. This remarkable fact is the starting point for the thorough analysis carried out in [8] to which the reader is referred for more details and the proofs of the results that are recalled without proof in the present study. Given its prominent role, the special loss value $\boldsymbol{L}_{(k+d)}(\boldsymbol{x}_N^*)$ is called *the Shortfall Threshold* and denoted by $\overline{\boldsymbol{L}}_N$ (as usual, the subscript N is used to recall that it is computed based on N observations),

$$\overline{\boldsymbol{L}}_N = \boldsymbol{L}_{(k+d)}(\boldsymbol{x}_N^*). \tag{9}$$

Corresponding to $\overline{\boldsymbol{L}}_N$, we introduce the Probability of Shortfall

$$\boldsymbol{PS}_N := \mathsf{P}\left\{\delta \in \Delta : \boldsymbol{L}(\boldsymbol{x}_N^*, \delta) > \overline{\boldsymbol{L}}_N\right\}. \tag{10}$$

This is the probability that a new $\delta \in \Delta$ drawn independently of the sample $(\delta^{(1)}, \delta^{(2)}, \cdots, \delta^{(N)})$ incurs a loss whose value is greater than the shortfall threshold at the solution of Problem (6). Since $\boldsymbol{x}_N^*$ depends on the random sample $(\delta^{(1)}, \delta^{(2)}, \cdots, \delta^{(N)})$, so does $\boldsymbol{PS}_N$, which is thus a random variable on $(\Delta^N, \mathcal{F}^N, \mathsf{P}^N)$. The following main theorem of [8] shows that the cumulative distribution function of $\boldsymbol{PS}_N$ is the same for all the problems that share the same parameters k and d, irrespective of the probability P.

Theorem 2.1 *$\boldsymbol{PS}_N$ is distributed as a Beta$(k+d, N+1-(k+d))$, thus its cumulative distribution function is*

$$\begin{aligned}\mathsf{P}^N\left\{\boldsymbol{PS}_N \le \varepsilon\right\} &= \int_0^\varepsilon (k+d)\binom{N}{k+d}\, p^{k+d-1}(1-p)^{N-k-d} dp \\ &= 1 - \sum_{i=0}^{k+d-1}\binom{N}{i}\varepsilon^i(1-\varepsilon)^{N-i}.\end{aligned} \tag{11}$$

$*$

Two corollaries are also proved in [8].

Corollary 2.2 *It holds that*

$$\mathsf{P}^{N+1}\left\{\boldsymbol{L}(\boldsymbol{x}_N^*, \delta^{(N+1)}) > \overline{\boldsymbol{L}}_N\right\} = \mathbb{E}\left\{\boldsymbol{PS}_N\right\} = \frac{k+d}{N+1}. \tag{12}$$

$*$

Corollary 2.3 *If $N \to \infty$ and k is allowed to grow with N so that $\lim_{N\to\infty} \frac{k}{N} = \varepsilon$, then $\boldsymbol{PS}_N \to \varepsilon$ in the mean-square sense.* $*$

In the present contribution, we are also interested in the distribution of the probability of exceeding other empirical costs than the Shortfall Threshold (a situation which is not considered in [8]). For $j = k+d, k+d+1, \cdots, N$, define

$$\boldsymbol{PS}_N^{(j)} := \mathsf{P}\left\{\delta \in \Delta : \boldsymbol{L}(\boldsymbol{x}_N^*, \delta) > \boldsymbol{L}_{(j)}(\boldsymbol{x}_N^*)\right\}. \tag{13}$$

$\boldsymbol{PS}_N^{(j)}$ is called the j-th Probability of Shortfall. Note also that $\boldsymbol{PS}_N = \boldsymbol{PS}_N^{(k+d)}$. The following result can be established, *mutatis mutandis*, by resorting to reasonings akin to those given in [11].

Theorem 2.4 *For all $j = k+d, k+d+1, \cdots, N$, the random variable $\boldsymbol{PS}_N^{(j)}$, $j = k+d, k+d+1, \cdots, N$, is distributed as a Beta$(j, N+1-j)$, thus its cumulative distribution function is*

$$\begin{aligned}\mathsf{P}^N\left\{\boldsymbol{PS}_N^{(j)} \le \varepsilon\right\} &= \int_0^\varepsilon j\binom{N}{j}\, p^{j-1}(1-p)^{N-j} dp \\ &= 1 - \sum_{i=0}^{j-1}\binom{N}{i}\varepsilon^i(1-\varepsilon)^{N-i}\ .\end{aligned} \tag{14}$$

$*$

2.4 Useful Tools for CVaR Data-Driven Decision-Making

The results presented so far can be employed to compute confidence intervals that are guaranteed to include the probabilities of shortfall with very high confidence (Subsection 2.4.1). This is instrumental to constructing a *probability box* that is guaranteed to contain the cumulative distribution function of $\boldsymbol{L}(\boldsymbol{x}_N^*, \delta)$ with a high confidence (Subsection 2.4.2).

2.4.1 Confidence Intervals

Let β be a confidence parameter, normally set to a small value such as 10^{-4}. By using Theorem 2.4, a confidence interval $[a, b]$ for $\boldsymbol{PS}_N^{(j)}$ at level $1-\beta$ can be constructed. The interval $[a, b]$ is obtained by computing the values a and b such that

$$\mathsf{P}^N\left\{\boldsymbol{PS}_N^{(j)} < a\right\} = \mathsf{P}^N\left\{\boldsymbol{PS}_N^{(j)} > b\right\} = \frac{\beta}{2}. \tag{15}$$

Note that a and b depend only on the parameters j and N that fully determine the distribution of $\boldsymbol{PS}_N^{(j)}$. When N is high enough, this distribution concentrates around its mean, given in (12) for $j = k + d$, with a thin tail. Therefore, even with very small values of β, the confidence interval will be small and useful.

2.4.2 Probability Boxes

A probability box is defined by a lower-bounding function $\boldsymbol{\Lambda}(l)$ and an upper-bounding function $\boldsymbol{\Gamma}(l)$ such that the relationship

$$\boldsymbol{\Lambda}(l) \leq \mathsf{P}\{\delta \in \Delta : \boldsymbol{L}(\boldsymbol{x}_N^*, \delta) \leq l\} \leq \boldsymbol{\Gamma}(l), \quad \forall l \in \mathbb{R} \tag{16}$$

holds true with confidence $1-\beta$. $\boldsymbol{\Lambda}(l)$ and $\boldsymbol{\Gamma}(l)$ can be obtained as follows. By using Theorem 2.4, let us construct $N-(k+d)+1$ confidence intervals $[u_j, v_j]$ for the variables $1-\boldsymbol{PS}_N^{(j)}$, $j = k+d, k+d+1, \cdots, N$, such that

$$\mathsf{P}^N\left\{1-\boldsymbol{PS}_N^{(j)} < u_j\right\} = \mathsf{P}^N\left\{1-\boldsymbol{PS}_N^{(j)} > v_j\right\} = \frac{\beta/2}{N+1-(k+d)}. \tag{17}$$

Then, by exploiting the monotonicity of the cumulative distribution function, one can easily see that (16) is satisfied with confidence $1-\beta$ [‡] with $\boldsymbol{\Lambda}(l)$ and $\boldsymbol{\Gamma}(l)$ defined as follows

$$\boldsymbol{\Lambda}(l) = \begin{cases} u_{k+d}, & \text{if } l \geq \boldsymbol{L}_{(k+d)}(\boldsymbol{x}_N^*), \\ u_j, & \text{if } \boldsymbol{L}_{(j)}(\boldsymbol{x}_N^*) \leq l < \boldsymbol{L}_{(j-1)}(\boldsymbol{x}_N^*) \text{ and } j = k+d+1, \cdots, N, \\ 0, & \text{if } l < \boldsymbol{L}_{(N)}(\boldsymbol{x}_N^*); \end{cases} \tag{18}$$

$$\boldsymbol{\Gamma}(l) = \begin{cases} 1, & \text{if } l > \boldsymbol{L}_{(k+d)}(\boldsymbol{x}_N^*), \\ v_j, & \text{if } \boldsymbol{L}_{(j+1)}(\boldsymbol{x}_N^*) < l \leq \boldsymbol{L}_{(j)}(\boldsymbol{x}_N^*) \text{ and } j = k+d, \cdots, N-1, \\ v_N, & \text{if } l \leq \boldsymbol{L}_{(N)}(\boldsymbol{x}_N^*). \end{cases} \tag{19}$$

[‡] Note that the construction is based on requiring that all the $N+1-(k+d)$ confidence intervals $[u_j, v_j]$ include the value of $1-\boldsymbol{PS}_N^{(j)}$, an event that happens on a set of observations $\delta^{(1)}, \delta^{(2)}, \cdots, \delta^{(N)}$ of probability at least $1-2(N+1-(k+d)) \cdot \frac{\beta/2}{N+1-(k+d)} = 1-\beta$, which can be set small enough to guarantee "practical certainty".

2.5 Balancing Shortfall Threshold and Probability of Shortfall

So far, k and d have been considered as fixed parameters. In practice, however, these parameters are tuning knobs and the user may want to consider solutions obtained for various values of k (corresponding to different attitudes towards uncertainty) and with various degrees of freedom d (for example, in portfolio optimization, $d+1$ is the number of assets that the decision-maker considers for possible investments). Grounded on the theoretical results discussed in previous sections, one can look for a satisfactory choice of the parameters k and d with the following procedure: given N scenarios, one solves problems of the form (6) for different choices of k and d, say $(d^{(1)},k^{(1)}),(d^{(2)},k^{(2)}),\cdots,(d^{(r)},k^{(r)})$. Correspondingly, r different solutions $\boldsymbol{x}^*_{d^{(1)},k^{(1)}|N},\boldsymbol{x}^*_{d^{(2)},k^{(2)}|N},\cdots,\boldsymbol{x}^*_{d^{(r)},k^{(r)}|N}$ and r thresholds $\overline{\boldsymbol{L}}_{d^{(1)},k^{(1)}|N},\overline{\boldsymbol{L}}_{d^{(2)},k^{(2)}|N},\cdots,\overline{\boldsymbol{L}}_{d^{(r)},k^{(r)}|N}$ are obtained. Denoting by $\boldsymbol{PS}_{d^{(i)},k^{(i)}|N}$ the Probability of Shortfall for each threshold $\overline{\boldsymbol{L}}_{d^{(i)},k^{(i)}|N}$, with the corresponding confidence interval $[a_{d^{(i)},k^{(i)}|N},b_{d^{(i)},k^{(i)}|N}]$, an application of the probability union bound yields

$$\begin{aligned}&\mathsf{P}^N\left\{\boldsymbol{PS}_{d^{(1)},k^{(1)}|N}\notin[a_{d^{(1)},k^{(1)}|N},b_{d^{(1)},k^{(1)}|N}]\text{ or}\cdots\text{ or }\boldsymbol{PS}_{d^{(r)},k^{(r)}|N}\notin[a_{d^{(r)},k^{(r)}|N},b_{d^{(r)},k^{(r)}|N}]\right\}\\&\leq\sum_{i=1}^{r}\mathsf{P}^N\left\{\boldsymbol{PS}_{d^{(i)},k^{(i)}|N}\notin[a_{d^{(i)},k^{(i)}|N},b_{d^{(i)},k^{(i)}|N}]\right\}=\sum_{i=1}^{r}\beta=r\beta.\end{aligned}\qquad(20)$$

Even when r is large, one can choose β small enough so that the overall confidence $1-r\beta$ is close to 1. For example, if $\beta=10^{-4}$, then even with 100 different choices of k and d the intervals for the Probability of Shortfall are simultaneously valid with confidence at least 99%. The Shortfall Thresholds $\overline{\boldsymbol{L}}_{d^{(1)},k^{(1)}|N},\overline{\boldsymbol{L}}_{d^{(2)},k^{(2)}|N},\cdots,\overline{\boldsymbol{L}}_{d^{(r)},k^{(r)}|N}$, the Empirical Conditional Values-at-Risk $\mathbf{CVaR}_{d^{(1)},k^{(1)}|N}$, $\mathbf{CVaR}_{d^{(2)},k^{(2)}|N}$, $\cdots$, $\mathbf{CVaR}_{d^{(r)},k^{(r)}|N}$, and the corresponding intervals for the Probability of Shortfall can then be plotted against the different choices of k and d in order to find the (k,d) pair that best fits the user's preferences. Moreover, when a more accurate analysis is required, the cumulative distribution functions of $\boldsymbol{L}(\boldsymbol{x}^*_{d,k|N},\cdot)$ for the various values of k and d can be compared by resorting to high-confidence probability boxes constructed as in Subsection 2.4.2.

3 A Study in Portfolio Optimization

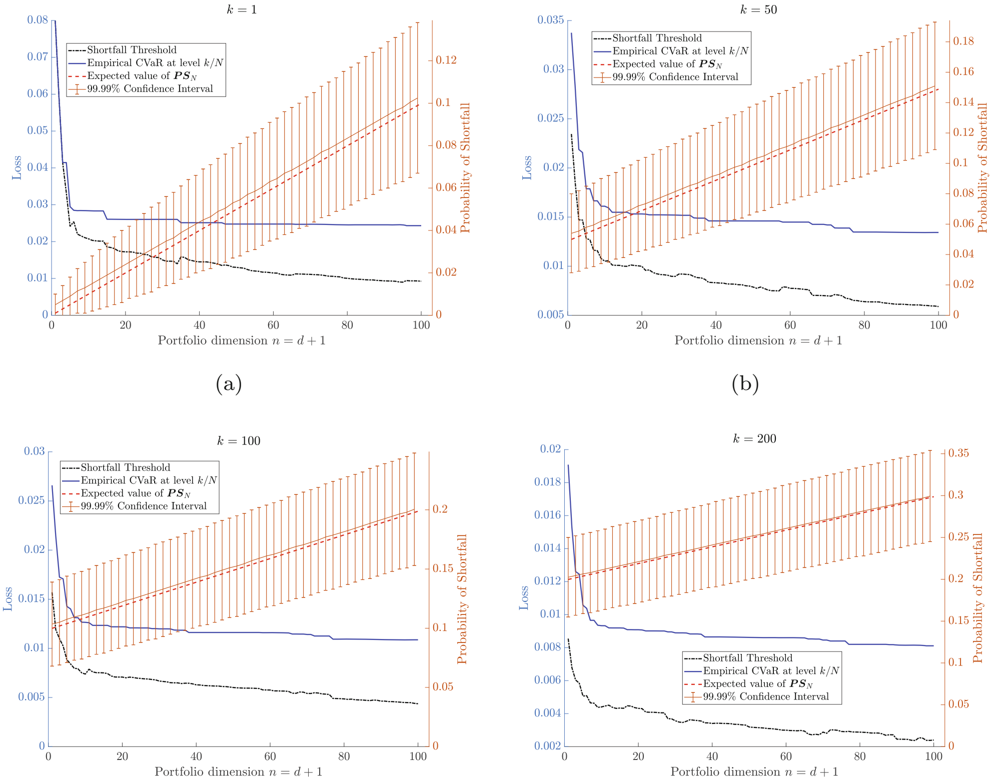
For our study, we collected the adjusted daily closing prices of the 300 assets with highest market capitalization as of March 1st, 2017 in S&P500 among those that reported a quote from March 1st, 2012 to March 1st, 2017 (data from finance.yahoo.com). As a result, we created a dataset with 1256 daily returns for each of the considered assets, starting from the most capitalized Apple Inc., Microsoft Corp., Exxon Mobil Corp. all the way through to Illinois Tool Works Inc.

3.1 Results

We considered portfolios with an increasing number of assets. The value $k+d$ (that influences both the mean and the dispersion of the distribution of $\boldsymbol{PS}_N$) increases with the number of assets $n=d+1$ and, as n grows, one observes the beneficial effect of diversification, while also experiencing a progressive increase of the probability of exceeding the Shortfall Threshold.

We started with a one-asset portfolio (only Apple Inc.) and evaluated the empirical CVaR and the empirical Shortfall Threshold. Then, we repeated the same procedure adding one asset at a time in order of decreasing capitalization[§].

Figure 1 shows the results obtained for four different values of k (1, 50, 100 and 200) and a number of assets n that ranged from 1 to 100 using data from the last $N = 1000$ days of the dataset. Notice that $k = 1$ corresponds to worst-case optimization and the other choices to the 5%, 10% and 20% empirical CVaR minimization, respectively. The plots display on the horizontal axis the portfolio dimension $n = d + 1$; the values of the CVaR and the Shortfall Threshold (blue solid and black dash-dotted lines) are read on the left vertical axis while the right vertical axis gives the values of the Probability of Shortfall.



(a)

(b)

(c)

(d)

Figure 1 Simulations for different values of k over $N = 1000$ days with portfolio increasing according to capitalization order

The first, although obvious, fact to be noticed is that the Empirical CVaR function is monotonically decreasing in n. In fact, when a new asset is included in the portfolio, two situations are possible:

[§] Some preliminary sorting of the assets is necessary to prevent the combinatorial proliferation of possible choices of n assets. Adding one set at a time in decreasing order of capitalization is just one sorting choice among many.

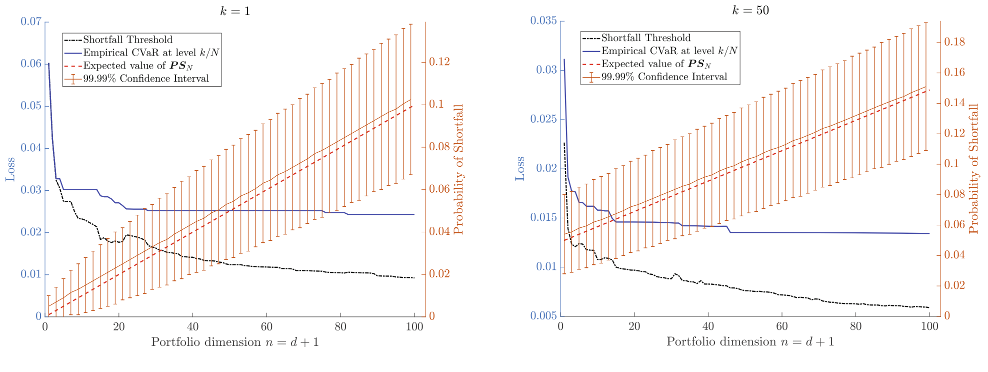
1. if the new asset produces an improvement in the cost function minimized in (6), then the portfolio is changed and a positive weight is allocated in the corresponding coordinate of $\boldsymbol{x}_N^*$;

2. if the addition of the new asset does not affect the optimization result, then a null weight is assigned to the corresponding coordinate of $\boldsymbol{x_N^*}$, the portfolio remains unchanged and so does the Empirical CVaR value.

The Shortfall Threshold function, instead, is not always monotonically decreasing, though it displays an overall decreasing trend (in general, including a new asset may result in a higher value of the Shortfall Threshold; certainly, however, in Case 2 above the added asset does not change the solution while $k + d$ increases so that the value of the Shortfall Threshold $\overline{\boldsymbol{L}}_N = \boldsymbol{L}_{(k+d)}(\boldsymbol{x}_N^*)$ decreases). We also observe a substantial increase of the gap between the Empirical CVaR and the Shortfall Threshold at the solution point as n increases.

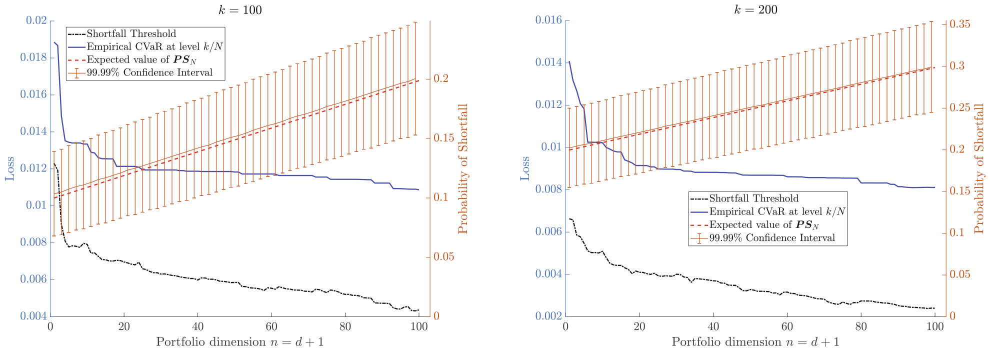
As n increases, the confidence interval for $\boldsymbol{PS}_N$ shifts up and widens. Hence, for increasing values of n, we obtain lower values of the Empirical CVaR (and of the Shortfall Threshold $\overline{\boldsymbol{L}}_N$) but the guarantees on the Probability of Shortfall worsen. In our experiments we adopted $\beta = 10^{-4}$ and, since the number of assets in the portfolio ranged from 1 to 100, the number r of solutions that we computed (one for each choice of assets) is 100. Therefore, the entire plot is guaranteed with confidence $1 - r\beta = 99\%$ (see Subsection 2.5). This allows an investor to consider different values of n, and hence different diversification levels both in terms of Empirical CVaR, Shortfall Threshold and confidence interval for $\boldsymbol{PS}_N$.

For a more complete analysis, we performed another test with the same settings, but this time, instead of adding assets in decreasing order of capitalization, we just sorted them randomly. The outcome demonstrates that the results are not significantly affected by the particular choice of the ordering. The results are shown in Figure 2.



(a)

(b)



(c) (d)

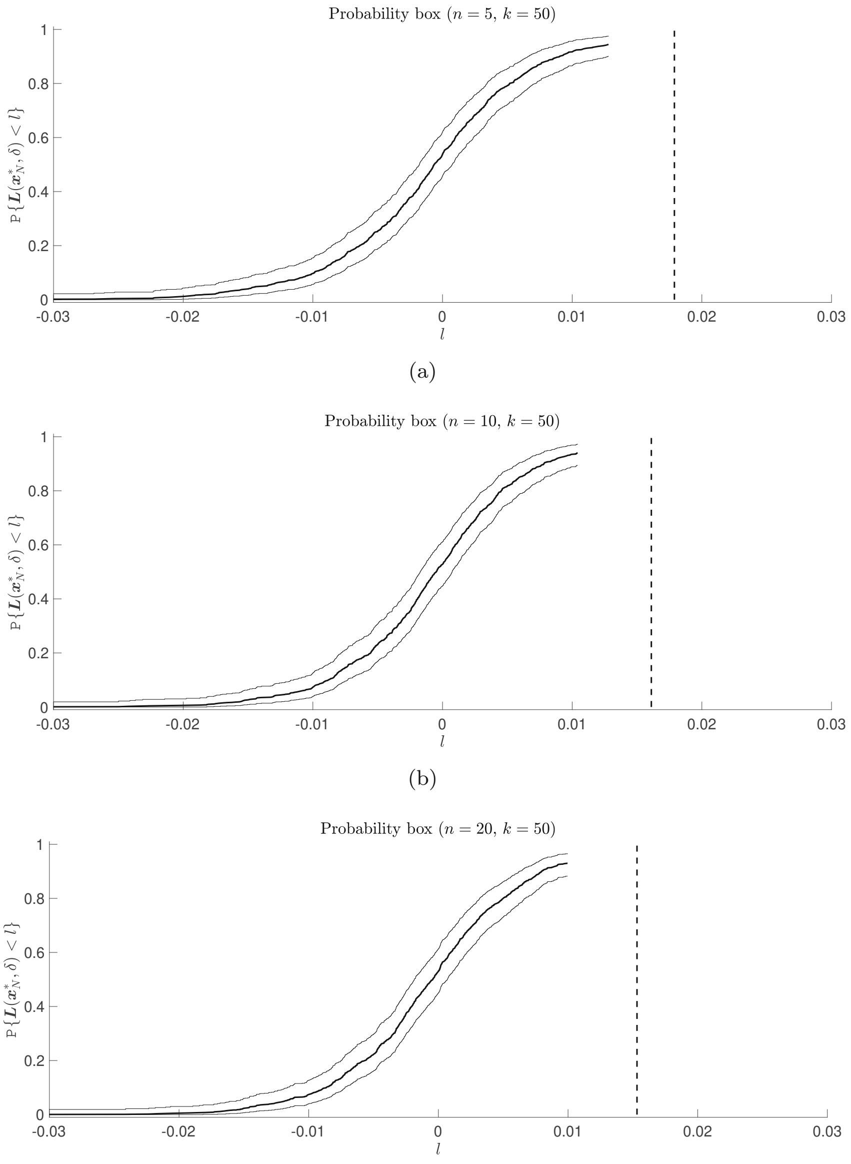
Figure 2 Simulations for different values of k over $N = 1000$ days with portfolio increasing according to a random order

3.2 Cumulative Distribution of the Loss

The results in this section are obtained by using $N = 1000$ scenarios from the market (as in the other simulations) and by optimizing the Empirical CVaR at 5% ($k = 50$) for portfolios of three different sizes $n = 5, 10, 20$ obtained from the capitalization ordering described before. At the point of minimum Empirical CVaR, i.e., at $\boldsymbol{x}_N^*$, we built the empirical cumulative distribution of the loss and constructed a probability box around it according to the procedure illustrated in Subsection 2.4.2; the probability box includes the true cumulative distribution function of the loss with confidence at least $1 - 10^{-4}$[¶].

It is worth noticing that the analysis offered in Subsection 3.1 was somehow defensive, as the main focus was put on the highest loss thresholds and on the probability of exceeding them. Referring to the probability boxes, one can inspect the probability of incurring losses as well as the probability of making a gain. When moving from panel (a) down to panels (b) and (c) in Figure 3, one also notes a decrease of dispersion in the portfolio loss (the cumulative distributions climb more rapidly); this implies a reduced volatility due to diversification, a phenomenon that plays an important role in sequential investments. Interestingly, while informative, the boxes are rigorously guaranteed distribution-free, that is, no assumptions are made on the underlying distribution by which the rates of return are generated (such as the assumption of log-normality which is often advocated in investment studies).

[¶] With our choices of the parameters, we have that $\max\{N - k - d + 1\} = 946$; therefore, a probability box with confidence larger than $1 - 10^{-4}$ is obtained by using (18) and (19) with intervals $[u_j, v_j]$ for $1 - \boldsymbol{PS}_N^{(j)}$ that are valid at level $1 - 10^{-7}$.



(a)

(b)

(c)

Figure 3 Empirical cumulative distribution of the loss at the solution point and 99.99% probability box

4 Concluding Remarks

In this paper, we have presented a set of theoretical tools that offer a well-principled environment for flexibility adjustment in empirical CVaR optimization. All results hold true without restrictive assumptions on the distributions by which observations are generated (distribution-free results). While this paper has by and large made reference to an illustrative application in portfolio selection, the tools here proposed are not application-dependent. In particular, it would be interesting to apply them to system identification and machine learning problems, where the trade-off between flexibility and generalization relates to the long-standing dilemma of balancing bias and variance effects.

References

[1] Vapnik V, *The Nature of Statistical Learning Theory*, Springer Science & Business Media, New York, 2013.

[2] Calafiore G and Campi M C, Uncertain convex programs: Randomized solutions and confidence levels, *Mathematical Programming*, 2005, **102**(1): 25–46.

[3] Calafiore G C and Campi M C, The scenario approach to robust control design, *IEEE Transactions on Automatic Control*, 2006, **51**(5): 742–753.

[4] Campi M and Garatti S, The exact feasibility of randomized solutions of uncertain convex programs, *SIAM Journal on Optimization*, 2008, **19**(3): 1211–1230.

[5] Campi M C and Garatti S, A sampling-and-discarding approach to chance-constrained optimization: Feasibility and optimality, *Journal of Optimization Theory and Applications*, 2011, **148**(2): 257–280.

[6] Ramponi F A, Consistency of the scenario approach, *SIAM Journal on Optimization*, 2018, **28**(1): 135–162.

[7] Garatti S, Campi M, and Carè A, On a class of interval predictor models with universal reliability, *Automatica*, 2019, **110**: 108542.

[8] Ramponi F A and Campi M C, Expected shortfall: Heuristics and certificates, *European Journal of Operational Research*, 2018, **267**(3): 1003–1013.

[9] Campi M C and Garatti S, Wait-and-judge scenario optimization, *Mathematical Programming*, 2018, **167**(1): 155–189.

[10] Carè A, Garatti S, and Campi M C, The wait-and-judge scenario approach applied to antenna array design, *Computational Management Science*, 2019, **16**: 481–499.

[11] Carè A, Garatti S, and Campi M C, Scenario min-max optimization and the risk of empirical costs, *SIAM Journal on Optimization*, 2015, **25**(4): 2061–2080.

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

[13] Ben-Tal A and Teboulle M, An old-new concept of convex risk measures: The optimized certainty equivalent, *Mathematical Finance*, 2007, **17**(3): 449–476.

[14] Mansini R, Ogryczak W, and Speranza M G, Conditional value at risk and related linear programming models for portfolio optimization, *Annals of Operations Research*, 2007, **152**(1): 227–256.

[15] Artzner P, Delbaen F, Eber J M, and Heath D, Coherent measures of risk, *Mathematical Finance*, 1999, **9**(3): 203–228.

[16] Pflug G C, Some remarks on the Value-at-Risk and the Conditional Value-at-Risk, *Probabilistic Constrained Optimization: Methodology and Applications* (Ed. by Uryasev S P), Boston, MA: Springer US, 2000, 272–281.

[17] Gotoh J Y and Takeda A, *CVaR Minimizations in Support Vector Machines*, John Wiley & Sons,

Ltd, 2016, 233–265.

[18] Chan T C, Mahmoudzadeh H, and Purdie T G, A robust-CVaR optimization approach with application to breast cancer therapy, *European Journal of Operational Research*, 2014, **238**(3): 876–885.

[19] Nasir H A, Carè A, and Weyer E, A randomised approach to flood control using Value-at-Risk, 2015 54*th IEEE Conference on Decision and Control* (*CDC*), 2015, 3939–3944.

[20] Filippi C, Guastaroba G, and Speranza M, Conditional value-at-risk beyond finance: A survey, *International Transactions in Operational Research*, 2020, **27**(3): 1277–1319.

[21] Hull J C, *Options, Futures and Other Derivatives*, Pearson, 2019.

[22] David H A and Nagaraja H N, *Order Statistics*, Wiley, New York, 2003.

[23] Carè A, Garatti S, and Campi M C, A coverage theory for least squares, *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*), 2017, **79**(5): 1367–1389.

Appendix

A Sufficient Condition for Non-Degeneracy in Portfolio Optimization

In portfolio optimization, Assumption 4 is implied by the following condition

Condition A.1 $\forall v \in \mathbb{R}^{d+1}$, $c \in \mathbb{R}$, v and c not both zeros in their respective spaces, it holds that $\mathsf{P}\{v^{\mathrm{T}}\delta = c\} = 0$. $*$

This condition is simpler than Assumption 4 because it involves only the realization of a single outcome δ. In what follows, we prove that Condition A.1 implies Assumption 4.

We will use the symbol $1\{\cdot\}$ to denote the indicator function: $1\{\text{statement}\}$ is equal to 1 when the statement is true and is 0 otherwise. Let D denote the matrix whose columns are $\delta^{(1)}, \delta^{(2)}, \cdots, \delta^{(d+1)}$. We need the following lemma.

Lemma A.1 *Under Condition* A.1, $\mathsf{P}^{d+1}\{D \ \ \textit{is invertible}\} = 1$.

Proof Condition A.1 immediately implies that $\mathsf{P}\{\delta = 0\} = 0$, so that

$$\mathsf{P}^{d+1}\{\delta^{(1)}, \delta^{(2)}, \cdots, \delta^{(d+1)} \ \ \text{are all nonzero vectors}\} = 1.$$

Denote by W a generic nonzero matrix of dimension $(d+1) \times d$ whose columns are the vectors $w_1, w_2, \cdots, w_d$. Let $\mathrm{colsp}(W)$ be the column space of W. Let $p(\cdot)$ be any vector function of W such that $p(W)$ is a non-zero vector in $\mathbb{R}^{d+1}$ and $p(W)^{\mathrm{T}} w_i = 0$ for all $i = 1, 2, \cdots, d$ (i.e., $p(W)$ is a vector in the orthogonal complement of $\mathrm{colsp}(W)$). Finally, denote by $D^{(i)}$ the matrix whose columns are the d vectors among $\delta^{(1)}, \delta^{(2)}, \cdots, \delta^{(d+1)}$ that are not $\delta^{(i)}$.

It holds that $\mathsf{P}^{d+1}\{D \text{ is singular}\} = \mathsf{P}^{d+1}\{\exists i \in \{1, 2, \cdots, d+1\} : \delta^{(i)} \in \mathrm{colsp}(D^{(i)})\} \leq \sum_{i=1}^{d+1} \mathsf{P}^{d+1}\{\delta^{(i)} \in \mathrm{colsp}(D^{(i)})\} \leq \sum_{i=1}^{d+1} \mathsf{P}^{d+1}\{{\delta^{(i)}}^{\mathrm{T}} p(D^{(i)}) = 0\}$. The Lemma is proven by noting that $\mathsf{P}^{d+1}\{{\delta^{(i)}}^{\mathrm{T}} p(D^{(i)}) = 0\} = \mathbb{E}_{D^{(i)} \in \Delta^d}\left[\mathbb{E}_{\delta^{(i)} \in \Delta}[1\{{\delta^{(i)}}^{\mathrm{T}} p(D^{(i)}) = 0\}]\right]$, and that the argument of the external expectation is equal to zero by Condition A.1. ▌

From now on, we assume that P satisfies Condition A.1 and hence that D is invertible by Lemma A.1 (the zero-probability event where this does not happen is unimportant and here neglected). When the event (8) is true, there exists a nonzero z such that $z^{\mathrm{T}}\delta^{(1)} = z^{\mathrm{T}}\delta^{(2)} = \cdots = z^{\mathrm{T}}\delta^{(d+1)} = z^{\mathrm{T}}\delta^{(d+2)}$. Then, it holds that $z^{\mathrm{T}}(\delta^{(i)} - \delta^{(d+2)}) = 0$, $i = 1, 2, \cdots, d+1$, which implies that the $d+1$ vectors $(\delta^{(i)}-\delta^{(d+2)})$, $i=1,2,\cdots,d+1$, are not linearly independent. This in turn entails that there exists a nonzero vector $y=[y_1,y_2,\cdots,y_{d+1}]^{\mathrm{T}}$ such that $\sum_{i=1}^{d+1} y_i(\delta^{(i)}-\delta^{(d+2)})=0$. Note that if $\sum_{i=1}^{d+1} y_i=0$, then D would be singular, which is ruled out under the present conditions. Hence, $\delta^{(d+2)}$ can be written as

$$\delta^{(d+2)}=\sum_{i=1}^{d+1}\frac{y_i}{\sum_{j=1}^{d+1} y_j}\delta^{(i)}, \tag{21}$$

which can also be expressed as

$$\delta^{(d+2)}=D\overline{y}, \tag{22}$$

where $\overline{y}$ is a vector whose components sum to 1. Letting $u=[1,1,\cdots,1]^{\mathrm{T}}$ be the vector made of $d+1$ ones, we can define $\widehat{v}=(D^{\mathrm{T}})^{-1}u$. Multiplying both sides of the equation (22) by $\widehat{v}^{\mathrm{T}}$ gives the equality $\widehat{v}^{\mathrm{T}}\delta^{(d+2)}=1$. With this result in mind, we get

$$\begin{aligned}&\mathsf{P}^{d+2}\{\text{event (8)}\}\\ \le\ &\mathbb{E}_{\delta^{(1)},\cdots,\delta^{(d+1)}\in\Delta^{d+1}}\left[\mathbb{E}_{\delta^{(d+2)}\in\Delta}[1\{\widehat{v}^{\mathrm{T}}\delta^{(d+2)}=1\}]\right]\\ \le\ &\mathbb{E}_{\delta^{(1)},\cdots,\delta^{(d+1)}\in\Delta^{d+1}}\left[\sup_v \mathbb{E}_{\delta^{(d+2)}\in\Delta}[1\{v^{\mathrm{T}}\delta^{(d+2)}=1\}]\right]\\ =\ &\mathbb{E}_{\delta^{(1)},\cdots,\delta^{(d+1)}\in\Delta^{d+1}}\left[\sup_v \mathsf{P}\{v^{\mathrm{T}}\delta=1\}\right]=0,\end{aligned}$$

by an application of Condition A.1. ▮