Ann Oper Res
https://doi.org/10.1007/s10479-017-2732-8

S.I.: RISK MANAGEMENT APPROACHES IN ENGINEERING APPLICATIONS

CVaR distance between univariate probability distributions and approximation problems

Konstantin Pavlikov[1] · **Stan Uryasev**[2]

Abstract The paper defines new distances between univariate probability distributions, based on the concept of the CVaR norm. We consider the problem of approximation of a discrete distribution by some other discrete distribution. The approximating distribution has a smaller number of atoms than the original one. Such problems, for instance, must be solved for generation of scenarios in stochastic programming. The quality of the approximation is evaluated with new distances suggested in this paper. We use CVaR constraints to assure that the approximating distribution has tail characteristics similar to the target distribution. The numerical algorithm is based on two main steps: (i) optimal placement of positions of atoms of the approximating distribution with fixed probabilities; (ii) optimization of probabilities with fixed positions of atoms. These two steps are iterated to find both optimal atom positions and probabilities. Numerical experiments show high efficiency of the proposed algorithms, solved with convex and linear programming.

Keywords Scenario reduction · Distance minimization · Conditional Value-at-Risk · CVaR norm

This research has been supported by the AFOSR Grant FA9550-12-1-0427, "Design and Redesign of Engineering Systems".

✉ Stan Uryasev
uryasev@ufl.edu

Konstantin Pavlikov
kop@sam.sdu.dk

[1] Department of Business and Economics, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark

[2] Risk Management and Financial Engineering Lab, Department of Industrial and Systems Engineering, University of Florida, 303 Weil Hall, Gainesville, FL 32611, USA

Published online: 15 January 2018

1 Introduction

The paper considers the problem of approximation of one probability distribution by some other distribution. Decision science literature considers various approaches for approximation of one probability distribution by another simpler one. Typically, a continuous distribution is approximated by a discrete one with a small number of atoms. For instance, three point approximations have been extensively studied, see Keefer and Bodily (1983) and Keefer (1994).

A number of approximation approaches has been suggested, including the mean (or the median) bracket or the moment matching methods. In the bracket approach, e.g., Miller and Rice (1983) and Hammond and Bickel (2013), the support of the target distribution is divided into several brackets (not necessary equal in probability) and the mean or the median of every bracket is chosen to be a discrete representation of that part of the target distribution. Another approach is based on the idea that the approximation should match the moments of the original distribution. Matching of moments is important in computing value lotteries and their certain equivalents, Smith (1993). The idea is as follows: the value function can be well approximated by a polynomial (with degree m) of a random variable. Thus, if the random variable is approximated by a simpler discrete variable having the same m first moments, then the expected value function is also well approximated. The key result: it is possible to match the first $2m - 1$ moments of the target distribution by a discrete one with only m atoms, see Miller and Rice (1983) and Smith (1993). Moreover, when the original distribution is not specified completely, so fewer than $2m - 1$ moments of the original distribution are known, the ambiguity is resolved by entropy maximization, see Rosenblueth and Hong (1987).

While matching moments is important in certain applications, usually, it is of interest to accurately approximate the cumulative distribution function (cdf) of the target distribution. Statistical literature measures the discrepancy between distributions using their cumulative distribution functions. For instance, the Kolmogorov–Smirnov distance is very popular, with the corresponding goodness of fit test, see, Gibbons and Chakraborti (2011) and Feller (1948). This distance is equal to the maximum of the absolute difference between cdfs, which is rather conservative measure. Several other distances are proposed, for instance the Cramer–von Mises distance, which is based on the area under the weighted squared difference between two cdfs, see Darling (1957). Also, the Anderson–Darling distance is based on the squared difference between two cdfs with a different weight function, see Boos (1981). The Kantorovich–Rubinstein distance is popular in various applications. It is defined as the cost of an optimal transportation of the probability mass from one distribution to another one, see Villani (2009), however, in one dimensional case, it equals the area between two cdfs, Vallander (1973).

We defined in this paper a new family of Conditional Value-at-Risk (CVaR) distances between distributions, which extends the notion of the Kolmogorov–Smirnov distance. Definition of Conditional Value-at-Risk and description of properties can be found in Rockafellar and Uryasev (2000, 2002). We consider the approximation problem of a discrete distribution by some other discrete distribution with a smaller number of atoms. The objective is to approximate well the cdf of the target distribution by minimizing the new CVaR distance. We suggest several approximation algorithms. One of the algorithms splits the approximation problem in two subproblems: (i) finding location of atoms with fixed probabilities; (ii) optimizing probabilities with fixed location of atoms. We show that both problems can be solved with linear or convex programming. A special case of the CVaR distance corresponds to the Kantorovich–Rubinstein distance. For this case, two subproblems (i), (ii) are combined in an iterative procedure that consecutively finds locations of atoms and their probabilities.

An accurate tail approximation of the target distribution is of special importance in various risk management applications. For instance, in finance it is important to correctly represent large losses occuring with low probabilities. We assure accuracy of tail approximation by imposing constraints on CVaRs of right and/or left tail of the loss distribution. Compared to the standard CVaR optimization (Rockafellar and Uryasev 2000) where probabilities of loss outcomes are known and fixed, here we deal with optimization with respect to both atoms probabilities and their locations. To our knowledge, CVaR was not considered so far in such a context. We proved that the constraint on CVaR with respect to probabilities of atoms is concave and can be linearized.

Finally, although the current paper considers one dimension approximations, this is an important subproblem in higher dimensions, see Grigoriu (2009). The cited paper reduces high dimensional problem to one-dimension approximation of cdfs of marginal distributions, their moments, and correlation matrices.

2 Risk-measure-based distance between probability distributions

This section provides an informal introduction to the notion of a risk-measure-based distance between distributions. Frequently, the distance between probability distributions is based on their cdfs. The cdf-based Kolmogorov–Smirnov distance (also called the uniform metric) is popular in statistical literature (Gibbons and Chakraborti 2011) as well as in distribution approximations, Smith (1993). This distance equals to the supremum of absolute difference between two cumulative distribution functions. It may be too conservative for discontinuous distributions (such as discrete distributions) since it is focused on some specific datapoint. We suggest the distance between two cdfs that is less conservative than the Kolmogorov–Smirnov distance. For two discrete distributions, the new distance equals the average of several largest discrepancies between two cumulative distribution functions. In a similar fashion, the distance can be defined based on quantile functions, where a quantile function is the inverse of a cdf function.

Let us remind the notion of the risk measure. A risk measure $\mathcal{R}$ is a map from a space of random variables to $\mathbb{R}$ with "loss" orientation, i.e., small values are preferable to large values. In our study $\mathcal{R}$ is a so-called coherent risk measure, see Artzner et al. (1999). It satisfies the following axioms (the axioms are slightly different from Artzner et al. (1999), due to Rockafellar and Uryasev (2013)):

- A1. $\mathcal{R}(\xi) = C$ for constant random variables $\xi = C$ a.s.
- A2. $\mathcal{R}(\xi_1) \leq \mathcal{R}(\xi_2)$ for $\xi_1 \leq \xi_2$ a.s.
- A3. $\mathcal{R}(\xi_1 + \xi_2) \leq \mathcal{R}(\xi_1) + \mathcal{R}(\xi_2)$
- A4. $\mathcal{R}(\lambda\xi_1) = \lambda\mathcal{R}(\xi_1)$, for any $\lambda \in (0, +\infty)$

Let F and G be two probability distributions on the set $\mathcal{A} = [a, b] \subset \mathbb{R}$ with cdfs denoted by $F(x)$ and $G(x)$, $x \in \mathcal{A}$. For defining the risk-measure-based distance we use an auxiliary random variable ξ on $\mathcal{A}$ with distribution H and density function $h(x) > 0$, $\forall x \in int(\mathcal{A})$.

Definition 1 The risk-measure-based distance between cumulative distributions F and G on $\mathcal{A}$ is defined by:

$$d^H(F, G) = \mathcal{R}\big(|F(\xi) - G(\xi)|\big). \tag{1}$$

The function $d^H(F, G)$ satisfies the properties of a probability metric, defined, for instance, in Rachev et al. (2008), Chapter 3.

Proposition 2.1 *Let F, Z, G be probability distributions on $\mathcal{A} = [a, b]$. If H is a distribution with a density function $h(x) > 0$, $\forall x \in int(\mathcal{A})$, then:*

1. $d^H(F, G) \geq 0$
2. $d^H(F, G) = 0 \iff \mu\big(\{x : F(x) \neq G(x)\}\big) = 0$ *where μ denotes Lebesgue measure*
3. $d^H(F, G) = d^H(G, F)$
4. $d^H(F, Z) \leq d^H(F, G) + d^H(G, Z)$

Proof This Proposition is a special case of the Proposition in "Appendix A". □

Section 7 similarly defines the *risk-measure-based distance between quantile functions*. Both cdf- and quantile-based distances are special cases of the so-called *risk-measure-based distance between maximal monotone relations* considered in "Appendix A". This general distance definition uses the concept of maximal monotone relation (Rockafellar and Royset 2014). Risk-measure-based cdf and quantile distances lead to corresponding approximation problems for distributions.

This paper is focused on the following special case, when F and G are discrete distributions with finite supports and $\mathcal{R}$ is the CVaR risk measure, which is a coherent risk measure, see, for instance, Pflug (2000) and Rockafellar and Uryasev (2002). We call this distance *CVaR distance between distributions F and G*.

3 CVaR distance for univariate discrete distributions with finite domains

This section considers CVaR distance for discrete distributions. Let $\mathcal{A} = [a, b]$ be a bounded closed interval in $\mathbb{R}$. Let F and G be two distributions on $\mathcal{A}$. In other words, a random variable with the distribution F or G takes values in $\mathcal{A}$ with probability 1. Let distribution F be defined by a set of outcomes $\mathbf{x} = (x_1, \ldots, x_n)$, $x_i \in \mathcal{A}$ with probabilities $p = (p_1, \ldots, p_n)$ and distribution G defined by a set of outcomes $\mathbf{y} = (y_1, \ldots, y_m)$, $y_j \in \mathcal{A}$ with probabilities $q = (q_1, \ldots, q_m)$. We assume without loss of generality that components of $\mathbf{x}$ and $\mathbf{y}$ are ordered, i.e., $x_i < x_j$ and $y_i < y_j$ for $i < j$. Cumulative distribution functions $F(\cdot)$ and $G(\cdot)$ are defined as follows:

$$F(z) = \sum_{i=1}^{n} p_i \, \mathbb{1}_{z \geq x_i}, \tag{2}$$

$$G(z) = \sum_{i=1}^{m} q_i \, \mathbb{1}_{z \geq y_i}, \tag{3}$$

where $\mathbb{1}_{z \geq x_i} = 1$ if $z \geq x_i$ and 0, otherwise. Moreover, the auxiliary random variable ξ with the distribution H is assumed to be uniformly distributed on $\mathcal{A}$ throughout the rest of the paper, $H = U(\mathcal{A})$. By definition, CVaR with confidence level α of a continuous random variable X is defined as follows:

$$\mathrm{CVaR}_\alpha(X) = E\left(X \middle| X > F^{-1}(\alpha)\right), \tag{4}$$

where $F^{-1}(\alpha)$ is the α-quantile of the random variable X, defined by

$$F^{-1}(\alpha) = \inf\{l \in \mathbb{R} : \mathbb{P}(X > l) \leq 1 - \alpha\}. \tag{5}$$

In a general case for $\alpha \in [0, 1)$, CVaR can be defined (see, Rockafellar and Uryasev (2000)) as follows:

$$\mathrm{CVaR}_\alpha(X) = \min_c \left(c + \frac{1}{1-\alpha} E\,[X - c]^+\right). \tag{6}$$

To be simple, CVaR of a random variable is the average of a specified percentage of its largest outcomes, sometimes in the literature referred to as the β-mean (Ogryczak 2010); see some basic examples with discrete distributions in Pavlikov and Uryasev (2014). The following definition presents the concept of CVaR norm of a random variable, introduced and studied in Mafusalov and Uryasev (2016).

Definition 2 Let $\mathcal{L}$ be a random variable. The CVaR norm of random $\mathcal{L}$ is defined as the CVaR with parameter α of the absolute value of $\mathcal{L}$:

$$\langle\langle \mathcal{L} \rangle\rangle_\alpha = \mathrm{CVaR}_\alpha(|\mathcal{L}|). \tag{7}$$

With the above definition of norm, the distance between distributions F and G is defined as follows.

Definition 3 Let ξ be a uniform random variable on $\mathcal{A} = [a, b]$ and $\alpha \in [0, 1)$. The CVaR distance between F and G is defined as

$$d_\alpha^U(F, G) = \langle\langle F(\xi) - G(\xi) \rangle\rangle_\alpha. \tag{8}$$

Let $\mathbf{t} = \{\mathbf{x} \cup \mathbf{y}\}$ be the union of sets of outcomes $\mathbf{x}$ and $\mathbf{y}$, with $t_1 = \min\{\mathbf{x} \cup \mathbf{y}\}$ and $t_s = \max\{\mathbf{x} \cup \mathbf{y}\}$. Then, the discrete random variable $F(\xi) - G(\xi)$ takes the following values

$$d_k = F(t_k) - G(t_k), \quad k = 1, \ldots, s-1, \tag{9}$$

with probabilities

$$\mathbb{P}(d_k) = \frac{t_{k+1} - t_k}{|\mathcal{A}|}, \quad k = 1, \ldots, s-1. \tag{10}$$

Figure 1 illustrates the definition of CVaR distance. The family of CVaR distances defined by (8) includes the Kolmogorov–Smirnov distance as a special case.

Definition 4 The Kolmogorov–Smirnov distance between two distributions with cumulative distribution functions $F(\cdot)$ and $G(\cdot)$ is defined as follows:

$$d_{KS}(F, G) = \sup_z |F(z) - G(z)|. \tag{11}$$

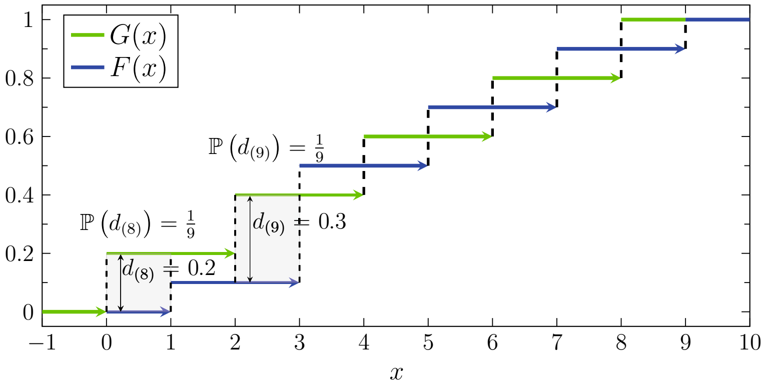


Fig. 1 CVaR distance between two discrete distributions, $\alpha = \frac{7}{9}$, $\mathcal{A} = [0, 9]$. The largest absolute difference between two cdfs, $d_{KS}(F, G) = d_{(9)} = 0.3$, has the probability $\frac{1}{9}$. The second largest absolute difference, $d_{(8)} = 0.2$ and also has the probability $\frac{1}{9}$. Therefore, $d_{7/9}^U(F, G) = \frac{0.3+0.2}{2} = 0.25$

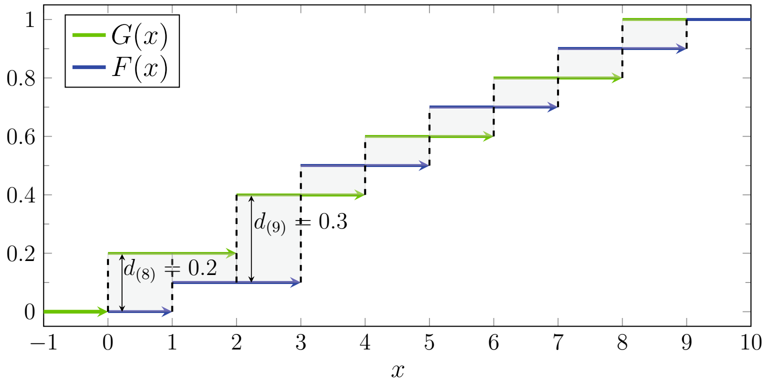


Fig. 2 The shaded area represents the Kantorovich–Rubinstein distance between F and G. The area scaled by the coefficient $\frac{1}{t_s - t_1} = \frac{1}{9}$ equals the average distance between F and G

The following remark establishes a connection between the Kolmogorov–Smirnov distance and the family of CVaR distances.

Remark 1 Definition 4 is a special case of Definition 3 when $\alpha \to 1$, i.e.,

$$d_{KS}(F,\, G) = \lim_{\alpha \to 1_-} d^U_\alpha(F,\, G) =: d^U_1(F,\, G).$$

Another special case is the CVaR distance with confidence level $\alpha = 0$, also called the average distance. The following definition explicitly presents the average distance between two discrete distributions.

Definition 5 The average distance between two distributions F and G, denoted by d^U_{AV}, is defined as follows:

$$d^U_{AV}(F,\, G) = \sum_{k=1}^{s-1} d_k \mathbb{P}(d_k), \tag{12}$$

with d_k and $\mathbb{P}(d_k)$ defined by (9) and (10).

Figure 2 illustrates the average distance between two discrete distributions.

Further we discuss relation between the average distance and another well-known distance, the Kantorovich–Rubinstein distance. The Kantorovich–Rubinstein distance between two discrete distributions is defined as follows.

Definition 6 (*Kantorovich–Rubinstein distance between two discrete distributions*) Define a transportation plan of transporting the probability mass of F to the distribution G, as follows:

$$\begin{aligned} w_{ij} &= \text{probability transported from outcome } y_j \text{ of } G \text{ to outcome } x_i \text{ of } F, \\ c_{ij} &= \text{transportation cost of unit of probability mass from } y_j \text{ to outcome } x_i. \end{aligned}$$

Here we assume that

$$c_{ij} = |x_i - y_j|, \quad i = 1, \ldots, n,\ j = 1, \ldots, m.$$

The Kantorovich–Rubinstein distance is defined as the optimal value of the following transportation problem:

$$d_K(F,\ G) = \min_{w_{ij}} \quad \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} w_{ij} \tag{13}$$

subject to

$$\sum_{i=1}^{n} w_{ij} = q_j, \quad j = 1, \ldots, m, \tag{14}$$

$$\sum_{j=1}^{m} w_{ij} = p_i, \quad i = 1, \ldots, n, \tag{15}$$

$$w_{ij} \geq 0, \quad i = 1, \ldots, n,\ j = 1, \ldots, m. \tag{16}$$

The following proposition establishes relation between d_{AV} and d_K .

Proposition 3.1 *Let distribution F be characterized by outcomes* $\mathbf{x}$ *and their probabilities* $\mathbf{p}$ *and the distribution G by outcomes* $\mathbf{y}$ *and their probabilities* $\mathbf{q}$. *Let* $\mathcal{A} = [t_1, t_s]$, *where* $t_1 = \min\{\mathbf{x} \cup \mathbf{y}\}$ *and* $t_s = \max\{\mathbf{x} \cup \mathbf{y}\}$. *Then, the cost of an optimal probability mass transportation plan,* (13)*, equals the scaled average distance,*

$$d_K(F,\ G) = (t_s - t_1) d_{AV}^U(F, G). \tag{17}$$

Proof As shown in Vallander (1973),

$$d_K(F,\ G) = \int_{\mathbb{R}} |F(z) - G(z)| dz = \int_{t_1}^{t_s} |F(z) - G(z)| dz.$$

Thus,

$$\int_{t_1}^{t_s} |F(z) - G(z)| dz = \sum_{i=1}^{s-1} d_k(t_{k+1} - t_k) = (t_s - t_1) d_{AV}(F,\ G).$$

□

4 Approximation of discrete distributions: distance minimization problem

This section defines the approximation problem of one discrete distribution by some other discrete distribution. We assume there exists a known reference distribution G with m outcomes, characterized by $(\mathbf{y}, \mathbf{q})$, and the goal is to find an approximation F, of a smaller size. In this section we assume that the outcomes of the approximating distribution, i.e., vector $\mathbf{x}$ is known. The objective is to find their probabilities $\mathbf{p}$ by minimizing the distance:

$$\min_{p_i} \ d_{\alpha}^U\ (F,\ G)\,, \tag{18}$$

$$\sum_{i=1}^{n} p_i = 1, \tag{19}$$

$$p_i \geq 0, \quad i = 1, \ldots, n. \tag{20}$$

The following proposition is needed for proving convexity of the problem (18)–(20).

Proposition 4.1 *Let $\mathcal{R}$ be a coherent risk measure. Then, $d^H(F,G)$ is a convex functional, i.e., with $\lambda \in (0,1)$, for any distributions F, G, $\widetilde{F}$, $\widetilde{G}$ on $\mathcal{A}$, the following holds:*

$$d^H(\lambda F + (1-\lambda)\widetilde{F}, \lambda G + (1-\lambda)\widetilde{G}) \le \lambda d^H(F,G) + (1-\lambda) d^H(\widetilde{F}, \widetilde{G}). \tag{21}$$

Proof Note that since $F(\cdot)$ associated with any distribution F is a nondecreasing function on $\mathcal{A}$, then the convex combination $\lambda F(\cdot) + (1-\lambda)\widetilde{F}(\cdot)$ is also a nondecreasing function on $\mathcal{A}$, and consequently $d^H(\lambda F + (1-\lambda)\widetilde{F}, \lambda G + (1-\lambda)\widetilde{G})$ is defined correctly. Using A2, A3, A4 axioms of coherent risk measures, we obtain:

$$d^H(\lambda F + (1-\lambda)\widetilde{F}, \lambda G + (1-\lambda)\widetilde{G}) = \tag{22}$$

$$\mathcal{R}(|\lambda F(\xi) + (1-\lambda)\widetilde{F}(\xi) - \lambda G(\xi) - (1-\lambda)\widetilde{G}(\xi)|) = \tag{23}$$

$$\mathcal{R}(|\lambda(F(\xi) - G(\xi)) + (1-\lambda)(\widetilde{F}(\xi) - (1-\lambda)\widetilde{G}(\xi))|) \overset{\text{A2}}{\le} \tag{24}$$

$$\mathcal{R}(|\lambda(F(\xi) - G(\xi))| + |(1-\lambda)(\widetilde{F}(\xi) - (1-\lambda)\widetilde{G}(\xi))|) \overset{\text{A3}}{\le} \tag{25}$$

$$\mathcal{R}(|\lambda(F(\xi) - G(\xi))|) + \mathcal{R}(|(1-\lambda)(\widetilde{F}(\xi) - (1-\lambda)\widetilde{G}(\xi))|) \overset{\text{A4}}{=} \tag{26}$$

$$\lambda\mathcal{R}(|F(\xi) - G(\xi)|) + (1-\lambda)\mathcal{R}(|\widetilde{F}(\xi) - \widetilde{G}(\xi)|) = \tag{27}$$

$$\lambda d^H(F,G) + (1-\lambda) d^H(\widetilde{F}, \widetilde{G}). \tag{28}$$

□

Proposition 4.1 implies the following corollary.

Corollary 4.1 *Let F be a discrete distribution on* $\mathbf{x}$ *with probabilities* $\mathbf{p}$*; G is a distribution on* $\mathbf{y}$ *with probabilities* $\mathbf{q}$. *Then $d_\alpha^U(F,G)$ is a convex function of variables* $(\mathbf{p}, \mathbf{q})$.

Corollary 4.1 implies that the problem (18)–(20) is convex in variables $\mathbf{p} = (p_1, \dots, p_n)$. Linearization of the problem (18)–(20) for various levels α can be done using standard approaches; we placed linearized formulations to "Appendix B".

5 CVaR distance minimization with cardinality constraint

The problem (18)–(20) in the previous section assumes that the outcomes $\mathbf{x}$ of the approximating distribution F are known. One way to relax this assumption is to consider that the set of outcomes $\mathbf{x}$ is a subset of points with a specified cardinality from some known set of points. For instance, we can consider the set of points from the target distribution $\mathbf{y}$. Consider now the problem (18)–(20) with the constraint that at most r out of m atoms of approximating distribution G are used. In other words, the number of outcomes of F with positive probabilities is less or equal to r. This is the problem formulation:

$$\min_{p_i, r_i} \; d_\alpha^U(F,G) \tag{29}$$

subject to

$$\sum_{i=1}^{m} r_i \le r, \tag{30}$$

$$\sum_{i=1}^{m} p_i = 1, \tag{31}$$

$$p_i \leq r_i, \quad i = 1, \ldots, m, \tag{32}$$

$$p_i \geq 0, \quad i = 1, \ldots, m, \tag{33}$$

$$r_i \in \{0, 1\}, \quad i = 1, \ldots, m. \tag{34}$$

The complete mixed integer linear programming formulation of this problem can be found in "Appendix B".

6 CVaR as a function of probabilities of atoms

This section studies properties of CVaR function w.r.t. probabilities of atoms. Such properties are used to impose CVaR constraints assuring a good fit of distribution tails.

As noted in Mason and Schuenemeyer (1983), the Kolmogorov–Smirnov distance is often insensitive to differences between distributions in tails. This statement is valid for the CVaR distance as well. When we approximate the distribution for the risk management purposes, it is desirable to ensure that the right tail of an approximating distribution is as heavy as the tail of the target distribution. We will impose the constraint that CVaR of the approximating distribution is not less than the CVaR of the target distribution. Moreover, we can impose CVaR constraints with several confidence levels.

Consider a discrete distribution G on the set of outcomes $\mathbf{y} = (y_1, \ldots, y_m)$ with probabilities $\mathbf{q} = (q_1, \ldots, q_m)$. Let an approximating distribution F be located at outcomes $\mathbf{x} = (x_1, \ldots, x_n)$ with unknown probabilities $\mathbf{p} = (p_1, \ldots, p_n)$. When we fit the distributions, the first thing that comes to mind is to match important characteristics of the distribution. One possibility is to impose a set of equality constraints on CVaRs. i.e., CVaRs of X should be equal to CVaRs of Y with several confidence levels:

$$\mathrm{CVaR}_{\alpha_i}(X) = \mathrm{CVaR}_{\alpha_i}(Y), \quad i = 1, \ldots, I. \tag{35}$$

Remark 2 Constraints (35) lead to nonconvex optimization problems. Case Study (2017) located online, "PROBLEM4: problem_KSMavg_Cvar_equality" employs nonlinear programming solution methods to solve such optimization problems and reports some results. However, the methods considered in the case study do not provide global optimality guarantee. This paper deals with the following inequality constraints preserving convexity and resulting in provably optimal solutions.

This section proposes the following constraints:

$$\mathrm{CVaR}_{\alpha_i}(X) \geq \mathrm{CVaR}_{\alpha_i}(Y), \quad i = 1, \ldots, I. \tag{36}$$

Such constraints ensure that the approximating distribution has at least as heavy right tail as the original one. Similarly, we can impose constraints on the left tail by changing the sign of the random variables:

$$-\mathrm{CVaR}_{\alpha_i}(-X) \leq -\mathrm{CVaR}_{\alpha_i}(-Y), \quad i = 1, \ldots, I. \tag{37}$$

Since the distribution of Y is known, then $\mathrm{CVaR}_{\alpha_i}(Y)$ is just a constant for every α_i. Therefore, both groups of constraints, (37) and (36), are of the same type,

$$\mathrm{CVaR}_{\alpha}(X) \geq a. \tag{38}$$

Further, we study constraint (38). The function $f_\alpha(\mathbf{p})$ defines the CVaR with a confidence level α of a random variable on outcomes $\mathbf{x}$ with **variable** vector of probabilities $\mathbf{p}$.

Definition 7 CVaR function $f_\alpha(\cdot)$ of variables $\mathbf{p}$ is defined as follows:

$$f_\alpha(\mathbf{p}) = \min_c \left(c + \frac{1}{1-\alpha} \sum_{i=1}^{n} p_i \left[x_i - c\right]^+ \right), \tag{39}$$

where

$$\left[x_i - c\right]^+ = \begin{cases} x_i - c, & \text{if } x_i \geq c, \\ 0, & \text{otherwise.} \end{cases}$$

The constraint (38) is expressed with $f_\alpha(\cdot)$ as

$$f_\alpha(\mathbf{p}) \geq a. \tag{40}$$

The function $f_\alpha(\mathbf{p})$ can be linearized; here are two representations:

1. Solution of the primal problem (39):

$$f_\alpha(\mathbf{p}) = \min_{c,\, z_i} \left(c + \frac{1}{1-\alpha} \sum_{i=1}^{n} p_i z_i \right) \tag{41}$$

subject to

$$z_i \geq x_i - c, \quad i = 1, \ldots, n, \tag{42}$$

$$z_i \geq 0, \quad i = 1, \ldots, n. \tag{43}$$

2. Solution of the problem dual to (41)–(43):

$$f_\alpha(\mathbf{p}) = \max_{w_i} \sum_{i=1}^{n} w_i x_i \tag{44}$$

subject to

$$w_i \leq \frac{p_i}{1-\alpha}, \quad i = 1, \ldots, n, \tag{45}$$

$$\sum_{i=1}^{n} w_i = 1, \tag{46}$$

$$w_i \geq 0, \quad i = 1, \ldots, n. \tag{47}$$

The following proposition establishes the concavity of $f_\alpha(\mathbf{p})$.

Proposition 6.1 *$f_\alpha(\mathbf{p})$ is a concave function.*

Proof Minimum of linear functions w.r.t. $\mathbf{p}$ is a concave function. This statement follows from the general theorem about convexity of the pointwise supremum of a set of convex functions, see, Theorem 5.5 in Rockafellar (1970). □

Corollary 6.1 *The set* $\{\mathbf{p} : f_\alpha(\mathbf{p}) \geq a\}$ *is convex.*

With representation (44)–(47) and Corollary 6.1, the convex constraint constraint (40) can be equivalently presented with the set of linear inequalities:

$$\sum_{i=1}^{n} w_i x_i \geq a \tag{48}$$

$$w_i \leq \frac{p_i}{1-\alpha}, \quad i = 1, \ldots, n, \tag{49}$$

$$\sum_{i=1}^{n} w_i = 1, \tag{50}$$

$$w_i \geq 0, \quad i = 1, \ldots, n. \tag{51}$$

Finally, we note that multiple constraints for various confidence levels α can be added to the distance minimization problems in Sects. 4 and 5. The constraints can be added to describe both left and right tails of the fitting distribution.

7 CVaR distance between quantile functions

Section 3 defined distances between distributions based on their cdfs. Sections 4, 5 and 6 discussed approaches for finding optimal probabilities of atoms of approximating distribution. This section defines another distance using quantiles of distributions. The new distance allows building optimization problems for finding optimal points (locations of atoms) of approximating distribution, when probabilities of atoms are known.

The quantile function of a probability distribution with the cdf $F(\cdot)$ is defined as follows:

$$F^{-1}(z) = \inf \{l \in \mathbb{R} : F(l) \geq z\}. \tag{52}$$

In financial risk management, the quantity $F^{-1}(z)$ is known as the Value-at-Risk (VaR). For instance, the function $F^{-1}(z) - G^{-1}(z)$ can represent the difference between VaRs of two portfolios. The maximum absolute difference $\sup_z |F^{-1}(z) - G^{-1}(z)|$ can be used to measure the distance between distributions. The notation F^{-1} is used to point out that the quantile is the inverse function to a cdf $F(\cdot)$. $F^{-1}(\cdot)$ is a nondecreasing function on $\mathcal{A} = [0, 1]$. We provide the following definition of the CVaR distance between quantiles F^{-1} and G^{-1}.

Definition 8 Let ξ be a uniform random variable on [0, 1]. CVaR distance between quantile functions is defined as follows:

$$d_\alpha^U \left(F^{-1}, G^{-1}\right) = \langle\langle F^{-1}(\xi) - G^{-1}(\xi)\rangle\rangle_\alpha. \tag{53}$$

Further we will specialize Definition 8 for discrete distributions. The quantile of a discrete distribution is equal to:

$$F^{-1}(z) = \inf \left\{ l \in \mathbb{R} : \sum_{i=1}^{n} p_i \mathbb{1}_{l \geq x_i} \geq z \right\}. \tag{54}$$

The presented expression for the quantile is rather complicated; it involves an optimization problem with the variable l in the argument of the indicator function. Let us rewrite this expression. The vector of probabilities $\mathbf{p} = (p_1, \dots, p_n)$ corresponds to the vector of outcomes $\mathbf{x}$, where, without loss of generality, the following can be assumed:

$$x_1 \leq \dots \leq x_n.$$

Let us define the cumulative distribution vector for the discrete distribution F:

$$\mathbf{f} = (f_1, \dots, f_n) = \left(p_1, p_1 + p_2, \dots, \sum_{i=1}^{n} p_i \right). \tag{55}$$

With these notations, the quantile F^{-1} equals:

$$F^{-1}(z) = x_{i_z}, \tag{56}$$

$$i_z = \min_i \, i \; : \; f_i \geq z, \;\; 0 \leq z \leq 1. \tag{57}$$

With notations for distribution G similar to (55), (56), (57) and the notation $\{\gamma_1, \dots, \gamma_s\} = \left\{p_1, p_1 + p_2, \dots, \sum_{i=1}^{n} p_i\right\} \cup \left\{q_1, q_1 + q_2, \dots, \sum_{j=1}^{m} q_j\right\}$, the random variable $F^{-1}(\xi) - G^{-1}(\xi)$ takes values

$$f_k^{-1} - g_k^{-1} = x_{i_{\gamma_k}} - y_{j_{\gamma_k}}, \quad k = 1, \dots, s-1, \tag{58}$$

$$i_{\gamma_k} = \min_i \, i : f_i \geq \gamma_k, \quad k = 1, \dots, s-1, \tag{59}$$

$$j_{\gamma_k} = \min_j \, j : g_j \geq \gamma_k, \quad k = 1, \dots, s-1, \tag{60}$$

with probabilities

$$\mathbb{P}\left(f_k^{-1} - g_k^{-1}\right) = \gamma_{k+1} - \gamma_k, \quad k = 1, \dots, s-1.$$

Figure 3 illustrates the definition of CVaR distance between quantile functions. This distance is used to address the following problem. Suppose the probability distribution with m outcomes G is known (as well as the function G^{-1}), and we would like to approximate it by a distribution F with n outcomes, for which we assume probabilities to be known. For instance, the approximating distribution is a uniformly distributed probability distribution, i.e., $p_i = \dfrac{1}{n}$, $i = 1, \dots, n$. The problem of finding the optimal positions $(x_1, \dots, x_n)$ is formulated as follows:

$$\min_{x_i} \quad d_\alpha^U \left(F^{-1}, \, G^{-1}\right) \tag{61}$$

subject to

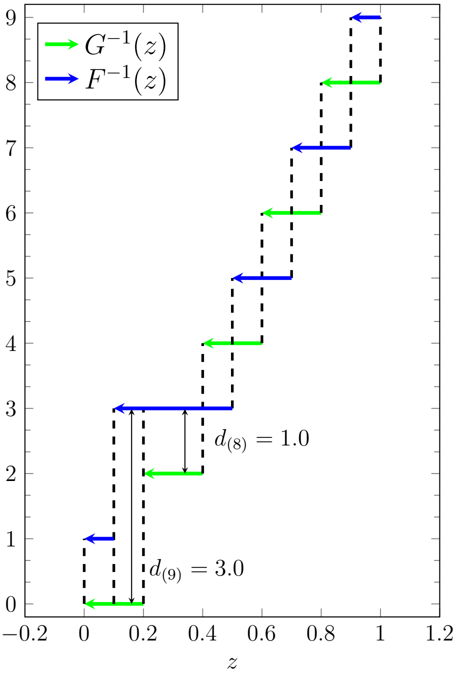
$$x_1 \leq \dots \leq x_n. \tag{62}$$

Convexity of the problem (61), (62) can be demonstrated following the proof of Proposition 4.1.

Corollary 7.1 *Let F be a discrete distribution on $\mathbf{x}$ with probabilities $\mathbf{p}$; G is a distribution on $\mathbf{y}$ with probabilities $\mathbf{q}$. Then $d_\alpha^U \left(F^{-1}, \, G^{-1}\right)$ is a convex function of variables $(\mathbf{x}, \, \mathbf{y})$.*

Proof Let F be a discrete probability distribution with ordered outcomes $\mathbf{x} = (x_1, \dots, x_n)$ and corresponding probabilities $\mathbf{p} = (p_1, \dots, p_n)$. Let $\widetilde{F}$ be a discrete probability distribution

Fig. 3 Illustration of the CVaR distance between quantile functions with confidence level $\alpha = \frac{7}{10}$. The largest absolute difference between two quantile functions, $d_{(9)} = 3.0$, has the probability of occurrence $\frac{1}{10}$. The second largest absolute difference, $d_{(8)} = 1.0$ and also has the probability of occurrence $\frac{2}{10}$. Thus, $d_{7/10}\left(F^{-1}, G^{-1}\right) = \frac{3.0\cdot 0.1+1.0\cdot 0.2}{0.3} \approx 1.67$



with ordered outcomes $\widetilde{\mathbf{x}} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$ and the same corresponding probabilities vector $\mathbf{p} = (p_1, \ldots, p_n)$. Similarly, G and $\widetilde{G}$ are two distributions on ordered $\mathbf{y} = (y_1, \ldots, y_m)$ and $\widetilde{\mathbf{y}} = (\widetilde{y}_1, \ldots, \widetilde{y}_m)$ having the same vector of probabilities $\mathbf{q} = (q_1, \ldots, q_m)$. With $\lambda \in (0, 1)$,

$$F_\lambda^{-1}(z) = \inf\left\{l \in \mathbb{R} : \sum_{i=1}^{n} p_i \mathbb{1}_{l \geq \lambda x_i + (1-\lambda)\widetilde{x}_i} \geq z\right\} = \lambda x_{i_z} + (1-\lambda)\widetilde{x}_{i_z} = \lambda F^{-1}(z) + (1-\lambda)\widetilde{F}^{-1}(z), \tag{63}$$

$$G_\lambda^{-1}(z) = \inf\left\{l \in \mathbb{R} : \sum_{j=1}^{m} q_j \mathbb{1}_{l \geq \lambda y_j + (1-\lambda)\widetilde{y}_j} \geq z\right\} = \lambda y_{j_z} + (1-\lambda)\widetilde{y}_{j_z} = \lambda G^{-1}(z) + (1-\lambda)\widetilde{G}^{-1}(z). \tag{64}$$

Then, by Proposition 4.1

$$d_\alpha^U\left(\lambda F^{-1} + (1-\lambda)\widetilde{F}^{-1},\ \lambda G^{-1} + (1-\lambda)\widetilde{G}^{-1}\right) \leq \lambda d_\alpha^U\left(F^{-1},\ G^{-1}\right) + (1-\lambda) d_\alpha^U\left(\widetilde{F}^{-1},\ \widetilde{G}^{-1}\right).$$

□

With the representation (58)–(60) of the random variable $F^{-1}(\xi) - G^{-1}(\xi)$, the linearization of the problem (61), (62) is straightforward; it is included in “Appendix B”.

8 Minimization of the Kantorovich–Rubinstein distance

It is easy to notice that for $\alpha = 0$ the distances $d_\alpha^U(F, G)$ and $d_\alpha^U\left(F^{-1}, G^{-1}\right)$ differ only by the scaling coefficient $(t_s - t_1)$. Moreover, for $\alpha = 0$, these distances measure the area between two cdf curves, and therefore they are equal to the Kantorovich–Rubinstein distance. Hence, these two measures can be used in one algorithm iterating two steps: (1) adjust probabilities of atoms; (2) adjust positions of atoms.

Let G be a known distribution characterized by $(\mathbf{y}, \mathbf{q})$. The approximating distribution F is required to have n atoms, i.e., the vectors of positions $\mathbf{x}$ and probabilities $\mathbf{p}$ have n components. Here is the outline for the iterative procedure for finding both $\mathbf{x}$ and $\mathbf{p}$:

- Step 0. Assign initial value for the vector $\mathbf{p}$, e.g., $\mathbf{p}^0 = \left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right)$, $k = 0$.
- Step 1. Find an optimal $\mathbf{x}^{k+1} = \left(x_1^{k+1}, \dots, x_n^{k+1}\right)$ by minimizing the quantile distance:
$$\min_{x_i^{k+1}} \; d_0^U \left(F^{-1}\left(\mathbf{p}^k, x_1^{k+1}, \dots, x_n^{k+1}\right), \; G^{-1}\right) \tag{65}$$
- Step 2. Find an optimal $\mathbf{p}^{k+1} = \left(p_1^{k+1}, \dots, p_n^{k+1}\right)$ by minimizing the cdf distance:
$$\min_{p_i^{k+1}} \; d_0^U \left(F\left(p_1^{k+1}, \dots, p_n^{k+1}, \mathbf{x}^{k+1}\right), \; G\right) \tag{66}$$
- $k := k + 1$, repeat Steps 1–2 until reduction in distance is getting smaller than threshold $\epsilon > 0$:
$$\left| d_0^U \left(F^{-1}\left(\mathbf{p}^{k+1}, \mathbf{x}^{k+1}\right), \; G^{-1}\right) - |\mathcal{A}| \, d_0^U \left(F\left(\mathbf{p}^k, \mathbf{x}^k\right), \; G\right) \right| < \epsilon. \tag{67}$$

This iterative procedure provides a series of approximations with non-increasing values of the Kantorovich–Rubinstein distance. The procedure generates a sequence of non-increasing numbers (distances) bounded from below (for instance, by 0), therefore the distances necessary converge to some value. The procedure does not necessary provide the global optimum for the distance minimization problem. The problem of approximating of one discrete distribution by some other discrete distribution is a combinatorial problem and here we outlined an efficient heuristic algorithm for this problem. Note that there exists a simple exact algorithm for minimum (Kantorovich–Rubinstein) distance approximation of a continuous distribution by a discrete one, see Kennan (2006).

9 Computational experiments

This section demonstrate the computational efficiency of the distance minimization algorithm described in the theoretical sections. The target probability distribution G is based on a real-life data set from the aerospace industry (recorded errors in the locations of fastener holes). The original dataset contains 8165 observations, with only 448 unique values. So, we assume that there are only 448 atoms with probabilities proportional to the number of observations at atoms. The dataset is further referred to as “Holes” dataset. Distance minimization problems were solved with two different solvers, Xpress (2014) and AORDA (2016): Portfolio Safeguard solver (PSG). For fair comparison, both solvers were set up to work using 1 thread. Computational results were obtained on a PC with Windows 8.1 × 64 operating system, Intel Core(TM) i5-4200M CPU 2.5 GHz, 6 GB RAM.

Table 1 presents computational results of the minimization problem (18)–(20) for a wide range of values of parameter α solved with linear (Xpress (2014)) and convex (AORDA (2016)) solvers. The problem (18)–(20) requires setting the outcomes $\mathbf{x}$ of the approximating probability distribution, which were selected uniformly across the range of the original

Table 1 Computational results of solving (18)–(20) problem

Dataset	m	n	α	Xpress		PSG	
				Objective	CPU time (s)	Objective	CPU time (s)
Holes	448	100	0.0	0.00253	0.14	0.00253	0.32
			0.1	0.00281	0.17	0.00281	0.37
			0.2	0.00316	0.17	0.00316	0.35
			0.3	0.00361	0.16	0.00361	0.38
			0.4	0.00419	0.19	0.00419	0.29
			0.5	0.00501	0.16	0.00501	0.29
			0.6	0.00619	0.18	0.00619	0.21
			0.7	0.00802	0.17	0.00802	0.14
			0.8	0.01056	0.15	0.01056	0.14
			0.9	0.01416	0.16	0.01416	0.11
			1.0	0.02995	0.06	0.02995	0.01

Table 2 Computational results of solving (61)–(62) problem

Dataset	m	n	α	Xpress	
				Objective	CPU time (s)
Holes	448	100	0.0	0.00058	0.03
			0.1	0.00065	0.04
			0.2	0.00073	0.04
			0.3	0.00083	0.03
			0.4	0.00097	0.03
			0.5	0.00116	0.04
			0.6	0.00137	0.03
			0.7	0.00166	0.04
			0.8	0.00224	0.03
			0.9	0.00399	0.04
			1.0	0.03300	0.02

distribution G. Similarly, Table 2 presents the computational results of approximating the original probability distribution by a uniform discrete distribution via the quantile distance minimization.

Table 3 presents computational results for the cardinality constrained version of the (18)–(20) problem. It is a challenging optimization problem, even for moderately sized instances. Mainly, this is due to the poor quality of the linear relaxation of the corresponding 0–1 problem formulation. Because of that, we compare the best objective values obtained by either of the solvers within a specified time limit. The optimality gaps for the considered instances and time limits are large, up to 95%, which suggests that both solvers essentially run as heuristics with little or no performance guarantee. One PSG minimization run of the problem with $r = 40$ is presented in Case Study (2017), see “PROBLEM2: problem_KSMavg_with_Cardinality”.

Finally, computational experiments of the iterative procedure for the Kantorovich–Rubinstein distance minimization (described in Sect. 8) are presented in Table 4. Experiments show that only few iterations are needed to achieve convergence of the procedure with a

Table 3 Computational experiments solving the cardinality constrained approximation problem (29)–(33)

Dataset	m	n	r	α	Xpress		PSG	
					Objective	CPU time (s)	Objective	CPU time (s)
Holes	448	448	10	0.0	0.0130	954.15	0.0140	955.58
				0.1	0.0156	739.80	0.0155	740.31
				0.2	0.0176	911.73	0.0183	912.12
				0.3	0.0197	936.78	0.0211	937.31
				0.4	0.0213	680.77	0.0237	681.04
				0.5	0.0264	1,504.77	0.0248	1,505.64
				0.6	0.0304	1,805.75	0.0278	1,806.03
				0.7	0.0346	1,244.74	0.0330	1,245.04
				0.8	0.0414	1,059.78	0.0483	1,060.46
				0.9	0.0494	1,997.78	0.0525	1,998.47
				1.0	0.0672	631.76	0.0763	632.42
Holes	448	448	20	0.0	0.0071	817.59	0.0078	818.21
				0.1	0.0081	580.76	0.0087	581.18
				0.2	0.0092	1,144.8	0.0106	1,145.61
				0.3	0.0103	1,698.75	0.0117	1,699.10
				0.4	0.0118	2,194.75	0.0111	2,195.01
				0.5	0.0138	1,776.74	0.0150	1,777.31
				0.6	0.0160	2,668.73	0.0150	2,669.80
				0.7	0.0187	2,498.7	0.0170	2,499.81
				0.8	0.0230	2,100.74	0.0242	2,101.60
				0.9	0.0247	1,524.74	0.0305	1,525.67
				1.0	0.0350	590.77	0.0360	591.79

The optimization problems were first run with AORDA software package. Solution time obtained by AORDA solver was used as an upper time limit for the FICO Xpress solver and the corresponding potentially suboptimal objective values are presented in the table

Table 4 Approximation of the dataset via minimization of the Kantorovich–Rubinstein distance

Dataset	m	n	Xpress		
			Objective	CPU time (s)	# of iterations
Holes	448	5	0.00551	0.91	5
		10	0.00321	1.71	7
		20	0.00186	3.06	9
		30	0.00125	4.00	10
		40	0.00106	1.37	5
		50	0.00079	3.92	9
		60	0.00066	2.72	7
		70	0.00049	6.55	11
		80	0.00045	4.87	9
		90	0.00038	9.98	13
		100	0.00034	10.53	13

Approximation is done both with respect to outcome positions and their probabilities

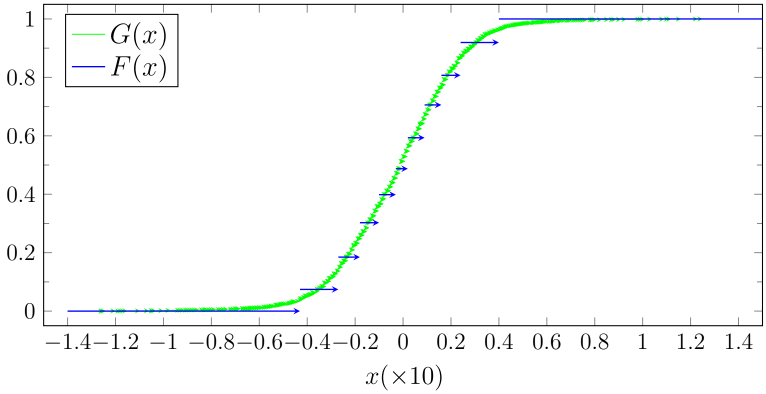


Fig. 4 An illustration of a result of the iterative Kantorovich–Rubinstein distance minimization procedure described in Sect. 8. Approximation is performed by a distribution with 10 outcomes

small, e.g., $\epsilon = 10^{-5}$, precision. Function "kantor" in PSG implements the algorithm. One minimization run of this function with $m = 40$ is presented in Case Study (2017), see "PROBLEM1: problem_Kantorovich_minimize". Figure 4 provides an illustration of the original dataset approximation by 10 discrete points using the iterative procedure. It is worth to note that contrary to the result of Kennan (2006), which states that the minimum Kantorovich–Rubinstein distance approximation of a continuous distribution is necessary uniform, the output of the procedure is not a uniform discrete distribution. The example also demonstrates that the right tail of the target distribution is heavier than the tail of the approximating distribution. Therefore, CVaR constraints discussed in Sect. 6 can be quite helpful in this case.

10 Conclusion

This paper described several approaches to the problem of approximation of one discrete distribution on the line with finite support by another one, potentially with a smaller number of outcomes. The approximation problem cam be split in two subproblems: (1) where to place the atoms of an approximating distribution, (2) which probabilities should be assigned to the atoms. These two problems are approached with distance measures between cumulative distribution functions and quantile functions, accordingly. An important special case is the Kantorovich–Rubinstein distance, which is equal to the area between functions (cdf or quantile) of two distributions. This fact allows to combine subproblems into an iterative procedure and minimize the Kantorovich–Rubinestein distance with respect to both atom locations and corresponding probabilities.

Some possible directions for future research may include the following. First, the cardinality constrained version of the approximation problem where atom positions of the approximation are selected from the set of original atom positions, may require a stronger problem formulation. Extension of the ideas behind the iterative Kantorovich–Rubinstein distance minimization procedure in higher dimensions can also be of interest. Finally, the introduced concept of CVaR distance as an extension of the Kolmogorov–Smirnov distance leads to the question whether it can be applied for hypothesis testing of equality of distributions, similar to the Kolmogorov–Smirnov test.

Appendix A: Risk-measure-based distance between maximal monotone relations

This section defines the notion of a risk-measure-based distance between two probability distributions in a formal way. It is convenient to formally define risk-measure-based distances based on *maximal monotone relations*, Rockafellar and Royset (2014).

A general cumulative distribution function can have points of discontinuity, however, if the corresponding jumps of such function are filled with vertical segments, we obtain an example of the maximal monotone relation. In a similar fashion, the quantile function of a probability distribution can generate a maximal monotone relation. We will consider further distances based on both cumulative distribution functions and quantile functions, so it is convenient to define the notion of distance in a more general fashion. First, we define the notion of a monotone relation on a set $\mathcal{A} \subseteq \mathbb{R}$.

Definition 9 (Rockafellar and Royset 2014). Let $\mathcal{A} = [a, b] \subseteq \mathbb{R}$, possibly unbounded closed interval. A set $\Gamma = \{(x,\ p) \subset \mathcal{A} \times \mathbb{R}\}$ is called a monotone relation on $\mathcal{A}$ if $\forall\ (x_1,\ p_1),\ (x_2,\ p_2) \in \Gamma$

$$(x_1 - x_2)(p_1 - p_2) \geq 0. \tag{A.1}$$

Set Γ is called a maximal monotone relation on $\mathcal{A}$ if there exists no monotone relation $\Gamma^{'} \neq \Gamma$ on $\mathcal{A}$, such that $\Gamma \subset \Gamma^{'}$.

Associated with a maximal monotone relation Γ on $\mathcal{A}$, the function $\Gamma(x),\ x \in \mathcal{A}$, is defined. An arbitrary monotone relation can clearly contain a vertical segment, therefore, let $\Gamma(x)$ be defined as

$$\Gamma(x) = \begin{cases} \inf\limits_{(x,\ p)\in\Gamma} \quad p, & \text{if} \quad x = b < +\infty, \\ \sup\limits_{(x,\ p)\in\Gamma} \quad p, & \text{otherwise}, \end{cases} \tag{A.2}$$

where b is the right point of the closed interval $\mathcal{A}$. Clearly, $\Gamma(x)$ is a nondecreasing function on $\mathcal{A}$.

Suppose F and G are two maximal monotone relations on $\mathcal{A}$ and we randomly pick a point $\xi \in \mathcal{A}$, so that the absolute difference between F and G becomes a random variable taking the value $|F(\xi) - G(\xi)|$. Specifically, we suppose there is an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the random variable ξ is a $\mathcal{F}-$measurable function from Ω to $\mathcal{A}$, $\xi : \Omega \to \mathcal{A}$. Let $\mathcal{A}$ be equipped with a Borel $\sigma-$algebra $\mathcal{B}$. Moreover, the auxiliary random variable ξ is supposed to have a probability distribution H, such that (i) it has a density function $h(x),\ x \in \mathcal{A}$ and (ii) $h(x) > 0$ for any $x \in int(\mathcal{A})$. The distance (discrepancy metric) between F and G will be defined using a risk measure to the random variable $|F(\xi) - G(\xi)|$.

A risk measure $\mathcal{R}$ is a map from a space of random variables into $\mathbb{R}$. In our study $\mathcal{R}$ belongs to a special class of risk measures, called coherent risk measures and defined in Artzner et al. (1999). To be coherent, a risk measure has to satisfy the following axioms (the axioms are in a slightly different from Artzner et al. (1999) form, due to Rockafellar and Uryasev (2013)):

- A1. $\mathcal{R}(\xi) = C$ for constant random variables $\xi = C$ a.s.,
- A2. $\mathcal{R}(\xi_1) \leq \mathcal{R}(\xi_2)$ for $\xi_1 \leq \xi_2$ a.s.,
- A3. $\mathcal{R}(\xi_1 + \xi_2) \leq \mathcal{R}(\xi_1) + \mathcal{R}(\xi_2)$,
- A4. $\mathcal{R}(\lambda\xi_1) = \lambda\mathcal{R}(\xi_1)$, for any $\lambda \in (0, +\infty)$.

Definition 10 The risk-measure-based distance between maximal monotone relations on $\mathcal{A}$, F and G, is defined through the corresponding functions $F(\cdot)$ and $G(\cdot)$, as follows:

$$d^H(F, G) = \mathcal{R}\big(|F(\xi) - G(\xi)|\big). \tag{A.3}$$

The function $d^H(F, G)$ satisfies the usual properties of a probability metric, discussed in, for instance, Rachev et al. (2008), Chapter 3.

Proposition A.1 *Let F, Z, G be maximal monotone relations on $\mathcal{A} = [a, b]$. If H is a distribution with a density function $h(x) > 0$, $\forall x \in int(\mathcal{A})$, the following properties hold:*

1. $d^H(F, G) \geq 0$
2. $d^H(F, G) = 0 \iff \mu\big(\{x : F(x) \neq G(x)\}\big) = 0$ *where μ denotes Lebesgue measure*
3. $d^H(F, G) = d^H(G, F)$
4. $d^H(F, Z) \leq d^H(F, G) + d^H(G, Z)$

Proof First, correctness of definition $|F(\xi) - G(\xi)|$ needs to be shown, i.e., that $|F(\xi) - G(\xi)|$ has to be a $\mathcal{F}-$ measurable function. It is sufficient to show that $F(\xi)$ is measurable, i.e., that the preimage of an open set in $\mathcal{A}$ is in $\mathcal{F}$. In order to see this sufficiency, note first that the sum or the difference of two measurable functions is measurable, see for instance McDonald and Weiss (1999), Chapter 3. Then, it is well-known that a preimage (with respect to a continuous function) of an open set is another open set, see McDonald and Weiss (1999), Chapter 2, therefore the absolute value function also preserves measurability. Thus, consider the measurability of the function $F(\cdot)$. The function $F(\cdot)$ is associated with a maximal monotone relation F, therefore it is nondecreasing and has at most countable number of points of discontinuity, cf. Rudin (1964) for example. Therefore, $F(\cdot)$ can be approximated by a sequence of continuous nondecreasing functions $F_n(\cdot)$ such that the sequence converges pointwise to F: $\forall x \in \mathcal{A}, \lim\limits_{n \to \infty} F_n(x) = F(x)$. By Theorem 4.5 in McDonald and Weiss (1999), the function $F(\cdot)$ is measurable.

Properties 1, 3, 4 are trivial and direct consequences of the axioms of coherent risk measures. The proof of Property 2 follows next.

1. We start from the $\Longrightarrow$ implication. Suppose $d^H(F, G) = 0$, in other words, $0 = \mathcal{R}(|F(\xi) - G(\xi)|) \leq \mathcal{R}(0)$, which implies by the property A2 of coherent risk measures that $\mathbb{P}(\omega : |F(\xi(\omega)) - G(\xi(\omega))| \leq 0) = \mathbb{P}(\omega : F(\xi(\omega)) = G(\xi(\omega))) = 1$. Thus, we obtain the following

$$\mathbb{P}\big(\omega : F(\xi(\omega)) \neq G(\xi(\omega))\big) = 0. \tag{A.4}$$

Let $\mathcal{A}^{'} \subset \mathcal{A}$ denote the image of ξ. Clearly, $\mu\left(\mathcal{A}^{'}\right) = \mu(\mathcal{A})$ because of the absolute continuity of the distribution of ξ. Then,

$$\mu\{x \in \mathcal{A} : F(x) \neq G(x)\} = \mu\left\{x \in \mathcal{A}^{'} : F(x) \neq G(x)\right\}. \tag{A.5}$$

Let $E = \left\{x \in \mathcal{A}^{'} : F(x) \neq G(x)\right\}$. Consider a sequence of $\{\epsilon_k > 0\}$, $\epsilon_k \to 0$, $k \to +\infty$ and let $E_k = \left\{x \in \mathcal{A}^{'} : F(x) \neq G(x),\ h(x) \geq \epsilon_k\right\}$. Clearly, $\bigcup\limits_{k=1}^{\infty} E_k = E$. Also,

$$0 = \mathbb{P}\big(\omega : F(\xi(\omega)) \neq G(\xi(\omega))\big) \geq \int_{E_k} h d\mu \geq \epsilon_k \int_{E_k} d\mu = \epsilon_k \mu(E_k) \implies \tag{A.6}$$

$$\mu\{E_k\} = 0, \quad k = 1, \ldots, +\infty. \tag{A.7}$$

Thus

$$\mu(E) = \mu\left(\bigcup_{k=1}^{\infty} E_k\right) \leq \sum_{k=1}^{\infty} \mu(E_k) = 0. \tag{A.8}$$

2. The implication $\Longleftarrow$ is more obvious. Let $E = \{x \in \mathcal{A} : F(x) \neq G(x)\}$. Then, due to ξ having the density function h,

$$\mathbb{P}\big(\omega : F(\xi(\omega)) \neq G(\xi(\omega))\big) = \int_E h d\mu = 0, \tag{A.9}$$

as an integral of the nonnegative function over the set of measure 0. □

Appendix B

Formulation 1 *The problem* (18)–(20) *for* $0 < \alpha < 1$ *can be reformulated as the following linear problem:*

$$\min_{c,\, p_i,\, z_k} \left(c + \frac{1}{1-\alpha} \sum_{k=1}^{s-1} \mathbb{P}(d_k) z_k \right) \tag{B.1}$$

subject to

$$z_k \geq \quad F(t_k) - G(t_k) - c, \quad k = 1, \ldots, s-1, \tag{B.2}$$

$$z_k \geq -F(t_k) + G(t_k) - c, \quad k = 1, \ldots, s-1, \tag{B.3}$$

$$F(t_k) = \sum_{i=1}^{n} p_i \mathbb{1}_{t_k \geq x_i}, \quad k = 1, \ldots, s-1, \tag{B.4}$$

$$\sum_{i=1}^{n} p_i = 1, \tag{B.5}$$

$$z_k \geq 0, \quad k = 1, \ldots, s-1, \tag{B.6}$$

$$p_i \geq 0, \quad i = 1, \ldots, n. \tag{B.7}$$

Formulation 2 *The problem* (18)–(20) *with* $\alpha = 1$ *can be reformulated as the following linear problem:*

$$\min_{a,\, p_i} \quad a \tag{B.8}$$

subject to

$$a \geq \quad F(t_k) - G(t_k), \quad k = 1, \ldots, s-1, \tag{B.9}$$

$$a \geq -F(t_k) + G(t_k), \quad k = 1, \ldots, s-1, \tag{B.10}$$

$$F(t_k) = \sum_{i=1}^{n} p_i \mathbb{1}_{t_k \geq x_i}, \quad k = 1, \ldots, s-1,$$

$$(B.5),\ (B.7). \tag{B.11}$$

Formulation 3 *The problem* (18)–(20) *with* $\alpha = 0$ *can be reformulated as the following linear problem:*

$$\min_{p_i,\, d_k} \quad \sum_{k=1}^{s-1} d_k \mathbb{P}(d_k) \tag{B.12}$$

subject to

$$d_k \geq \quad F(t_k) - G(t_k), \quad k = 1, \ldots, s-1, \tag{B.13}$$

$$d_k \geq -F(t_k) + G(t_k), \quad k = 1, \ldots, s-1, \tag{B.14}$$

$$F(t_k) = \sum_{i=1}^{n} p_i \, \mathbb{1}_{t_k \geq x_i}, \quad k = 1, \ldots, s-1,$$

$$(B.5),\ (B.7). \tag{B.15}$$

Formulation 4 *The problem* (29)–(33) *with* $\alpha \in (0, 1)$ *can be reformulated as the following linear problem:*

$$\min_{c,\, p_i,\, r_i,\, z_k} \left(c + \frac{1}{1-\alpha} \sum_{k=1}^{m-1} \mathbb{P}(d_k) z_k \right) \tag{B.16}$$

subject to

$$z_k \geq \quad F(y_k) - G(y_k) - c, \quad k = 1, \ldots, m-1, \tag{B.17}$$

$$z_k \geq -F(y_k) + G(y_k) - c, \quad k = 1, \ldots, m-1, \tag{B.18}$$

$$F(y_k) = \sum_{i=1}^{m} p_i \, \mathbb{1}_{y_k \geq y_i}, \quad k = 1, \ldots, m-1, \tag{B.19}$$

$$p_i \leq r_i, \quad i = 1, \ldots, m, \tag{B.20}$$

$$\sum_{i=1}^{m} r_i \leq r, \tag{B.21}$$

$$z_k \geq 0, \quad k = 1, \ldots, m-1, \tag{B.22}$$

$$r_i \in \{0, 1\}, \quad i = 1, \ldots, m,$$

$$(B.5),\ (B.7). \tag{B.23}$$

Corresponding reformulations for problem (29)–(33) *with* $\alpha = 1$ *and* $\alpha = 0$ *can be obtained similar to Formulations* 2 *and* 3.

Formulation 5 *The problem* (61)–(62) *with* $\alpha \in (0, 1)$ *for the minimization of CVaR distance between quantile functions can be reformulated as the following linear problem:*

$$\min_{c,\, x_i,\, z_k} \left(c + \frac{1}{1-\alpha} \sum_{k=1}^{s-1} \mathbb{P}(d_k) z_k \right) \tag{B.24}$$

subject to

$$z_k \geq \quad x_{i_{\gamma_k}} - y_{j_{\gamma_k}} - c, \quad k = 1, \ldots, s-1, \tag{B.25}$$

$$z_k \geq -x_{i_{\gamma_k}} + y_{j_{\gamma_k}} - c, \quad k = 1, \ldots, s-1, \tag{B.26}$$

$$z_k \geq 0, \quad k = 1, \ldots, s-1, \tag{B.27}$$

$$x_1 \leq \ldots \leq x_n. \tag{B.28}$$

Corresponding reformulations for problem (61)–(62) *with* $\alpha = 1$ *and* $\alpha = 0$ *can be obtained similar to Formulations* 2 *and* 3.

References

AORDA. (2016). Portfolio Safeguard Version 2.3. http://www.aorda.com/index.php/portfolio-safeguard/.

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203–228.

Boos, D. D. (1981). Minimum distance estimators for location and goodness of fit. *Journal of the American Statistical Association*, *76*(375), 663–670.

Case Study. (2017). Minimization of Kantorovich-Rubinstein distance between two distributions. http://www.ise.ufl.edu/uryasev/research/testproblems/advanced-statistics/minimize_kantorovich_distance/.

Darling, D. A. (1957). The Kolmogorov–Smirnov, Cramer–von Mises tests. *The Annals of Mathematical Statistics*, *28*, 823–838.

Feller, W. (1948). On the Kolmogorov–Smirnov limit theorems for empirical distributions. *The Annals of Mathematical Statistics*, *19*(2), 177–189.

Gibbons, J. D., & Chakraborti, S. (2011). Nonparametric Statistical Inference. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 977–979). Berlin: Springer. https://doi.org/10.1007/978-3-642-04898-2_420.

Grigoriu, M. (2009). Reduced order models for random functions. Application to stochastic problems. *Applied Mathematical Modelling*, *33*(1), 161–175.

Hammond, R. K., & Bickel, J. E. (2013). Reexamining discrete approximations to continuous distributions. *Decision Analysis*, *10*(1), 6–25.

Keefer, D. L. (1994). Certainty equivalents for three-point discrete-distribution approximations. *Management Science*, *40*(6), 760–773.

Keefer, D. L., & Bodily, S. E. (1983). Three-point approximations for continuous random variables. *Management Science*, *29*(5), 595–609.

Kennan, J. (2006). *A note on discrete approximations of continuous distributions*. Madison: University of Wisconsin.

Mafusalov, A., & Uryasev, S. (2016). CVaR (Superquantile) norm: Stochastic case. *European Journal of Operational Research*, *249*(1), 200–208.

Mason, D. M., & Schuenemeyer, J. H. (1983). A modified Kolmogorov–Smirnov test sensitive to tail alternatives. *The Annals of Statistics*, *11*(3), 933–946.

McDonald, J. N., & Weiss, N. A. (1999). A course in real analysis. Academic Press. https://books.google.dk/books?id=T-PUyB9YpqcC.

Miller, A. C. III., Rice, T. R. (1983). Discrete approximations of probability distributions. *Management Science, 29*(3), 352–362. https://doi.org/10.1287/mnsc.29.3.352.

Ogryczak, W. (2010). On robust solutions to multi-objective linear programs. *Multiple Criteria Decision Making*, *9*, 197–212.

Pavlikov, K., & Uryasev, S. (2014). CVaR norm and applications in optimization. *Optimization Letters*, *8*(7), 1999–2020.

Pflug, G. C. (2000). Some remarks on the Value-at-Risk and the Conditional Value-at-Risk. In S. P. Uryasev (Ed.), *Probabilistic constrained optimization: Methodology and applications* (pp. 272–281). Boston: Springer. https://doi.org/10.1007/978-1-4757-3150-7_15.

Rachev, S. T., Stoyanov, S. V., & Fabozzi, F. J. (2008). *Advanced stochastic models, risk assessment, and portfolio optimization: The ideal risk, uncertainty, and performance measures* (Vol. 149). Hoboken: Wiley.

Rockafellar, R. T. (1970). *Convex analysis* (Vol. 28). Princeton: Princeton University Press.

Rockafellar, R. T., & Royset, J. O. (2014). Random variables, monotone relations, and convex analysis. *Mathematical Programming*, *148*(1–2), 297–331.

Rockafellar, R. T., & Uryasev, S. (2000). Optimization of Conditional Value-at-Risk. *Journal of Risk*, *2*(3), 21–41.

Rockafellar, R. T., & Uryasev, S. (2002). Conditional Value-at-Risk for general loss distributions. *Journal of Banking and Finance*, *26*(7), 1443–1471.

Rockafellar, R. T., & Uryasev, S. (2013). The fundamental risk quadrangle in risk management, optimization and statistical estimation. *Surveys in Operations Research and Management Science*, *18*(1), 33–53.

Rosenblueth, E., & Hong, H. P. (1987). Maximum entropy and discretization of probability distributions. *Probabilistic Engineering Mechanics*, *2*(2), 58–63.

Rudin, W. (1964). *Principles of mathematical analysis* (Vol. 3). New York: McGraw-Hill.

Smith, J. E. (1993). Moment methods for decision analysis. *Management Science*, *39*(3), 340–358.

Vallander, S. S. (1973). Calculation of the Wasserstein distance between probability distributions on the line. *Teoriya Veroyatnostei i ee Primeneniya*, *18*(4), 784–786.

Villani, C. (2009). *Optimal transport: Old and new* (Vol. 338). Berlin: Springer.

Xpress, (2014). FICO™ Xpress Optimization Suite 7.8. http://www.fico.com.