ALSO-X and ALSO-X+: Better Convex Approximations for Chance Constrained Programs

Nan Jiang
Department of Industrial & Systems Engineering, Virginia Tech, Blacksburg, VA 24061, jnan97@vt.edu
Weijun Xie
Department of Industrial & Systems Engineering, Virginia Tech, Blacksburg, VA 24061, wxie@vt.edu

In a chance constrained program (CCP), decision-makers seek the best decision whose probability of violating the uncertainty constraints is within the prespecified risk level. As a CCP is often nonconvex and is difficult to solve to optimality, much effort has been devoted to developing convex inner approximations for a CCP, among which the conditional value-at-risk (CVaR) has been known to be the best for more than a decade. This paper studies and generalizes the ALSO-X, originally proposed by **A**hmed, **L**uedtke, **SO**ng, and **X**ie (2017), for solving a CCP. We first show that the ALSO-X resembles a bilevel optimization, where the upper-level problem is to find the best objective function value and enforce the feasibility of a CCP for a given decision from the lower-level problem, and the lower-level problem is to minimize the expectation of constraint violations subject to the upper bound of the objective function value provided by the upper-level problem. This interpretation motivates us to prove that when uncertain constraints are convex in the decision variables, ALSO-X always outperforms the CVaR approximation. We further show (i) sufficient conditions under which ALSO-X can recover an optimal solution to a CCP; (ii) an equivalent bilinear programming formulation of a CCP, inspiring us to enhance ALSO-X with a convergent alternating minimization method (ALSO-X+); (iii) an extension of ALSO-X and ALSO-X+ to distributionally robust chance constrained programs (DRCCPs) under $\infty-$Wasserstein ambiguity set. Our numerical study demonstrates the effectiveness of the proposed methods.

Key words: Chance Constraint; CVaR; Distributionally Robust; Bilievel Optimization

1. Introduction

Let us consider a chance constrained program (CCP) of the form

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \mathbb{P}\left\{ \tilde{\boldsymbol{\xi}} \colon g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \le 0 \right\} \ge 1-\varepsilon \right\}. \tag{1}$$

The goal of CCP (1) is to find a solution $\boldsymbol{x} \in \mathcal{X}$ that minimizes the objective $\boldsymbol{c}^\top \boldsymbol{x}$ and is subject to the uncertain constraints $g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \le 0$ satisfied with probability $1-\varepsilon$, where $\varepsilon \in (0,1)$ is a preset risk level. In this paper, we focus on the convex setting, i.e., throughout this paper, we make the following assumptions

- A1 Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where the probability measure $\mathbb{P}$ is defined on the measurable space $(\Omega, \mathcal{F})$ equipped with the sigma algebra $\mathcal{F}$, the random vector $\tilde{\boldsymbol{\xi}} : \Omega \to \Xi$ is a measurable mapping from Ω to $\mathbb{R}^m$ with support set $\Xi \subseteq \mathbb{R}^m$. Function $g(\boldsymbol{x}, \boldsymbol{\xi}) = \max_{i\in[I]} g_i(\boldsymbol{x}, \boldsymbol{\xi})$, where $g_i(\boldsymbol{x}, \boldsymbol{\xi}) \colon \mathbb{R}^n \times \Xi \to \mathbb{R}$ for all $i \in [I] := \{1, \ldots, I\}$ and $g_i(\boldsymbol{x}, \boldsymbol{\xi})$ is convex and lower semicontinuous in $\boldsymbol{x}$ for almost every $\boldsymbol{\xi} \in \Xi$;
- A2 Set $\mathcal{X}$ is nonempty, closed, convex, contained in a closed convex cone $\mathcal{C}$. That is, $\emptyset \neq \mathcal{X} \subseteq \mathcal{C}$, where $\mathcal{C}$ is a closed pointed convex cone; and
- A3 The feasible region of CCP (1) is nonempty and the objective cost vector $\boldsymbol{c} \in \mathrm{int}(\mathcal{C}^*) \cup \{\boldsymbol{0}\}$, where $\mathcal{C}^*$ is the dual cone of $\mathcal{C}$ and $\mathrm{int}(\cdot)$ denotes the interior of a set.

If $I = 1$, CCP (1) involves a single chance constraint, and otherwise, it contains a joint chance constraint. Note that Assumption A1 follows from existing chance constraint literature (see, e.g., Nemirovski and Shapiro 2007, Ahmed et al. 2017, Ahmed and Xie 2018) and the lower semicontinuity assumption in Assumption A1 and the closedness of set $\mathcal{X}$ in Assumption A2 together guarantee that the feasible region of CCP (1) is closed (see Proposition 11 in Appendix D), while Assumption A2 is quite standard for convex analysis (see, e.g., section 5 in Nemirovski 2001), and Assumption A3 guarantees that there exists an optimal solution in CCP (1) (see also the discussions in Xie and Ahmed 2020), for the sake of simplicity. It is worthy of mentioning that if Assumption A2 does not hold, for example, if set $\mathcal{X}$ is mixed-integer, then the main result that ALSO-X is better than CVaR approximation does not hold (see Example 2). Thus, the convexity assumption in Assumption A2 is crucial to this key result.

1.1. Relevant Literature

Since its first appearance to tackle uncertain constraints in the decision-making problems (Charnes and Cooper 1963, Charnes et al. 1958), CCPs have been studied and applied in many areas. For example, Pagnoncelli et al. (2009) considered a portfolio selection problem, where the decision-makers plan to achieve the targeted return rate with high probability. Chance constraints have also been employed to ensure a high level of service in transportation assignment problems (Dentcheva et al. 2000) or facility location problems (Lejeune and Margot 2016). In power systems (see, e.g., Bienstock et al. 2014, Shiina 1999, Xie and Ahmed 2017, Zhang et al. 2016), the decision-makers would like to restrict the probability of capacity violations of transmission lines within a small risk level. Deng and Shen (2016) studied a scheduling problem in healthcare, where planners want to have a low level of overtime servers. Interested readers are referred to the work by Ahmed and Shapiro (2008) for more CCP applications. Albeit important, a CCP encounters two main difficulties — its feasible region is often nonconvex, and checking the feasibility of a CCP for a given solution, in general, is challenging.

To address the aforementioned difficulties, there are several approaches proposed in the literature to solve CCP (1). One method is to investigate the conditions where the feasible region in CCP (1) is convex. For example, as demonstrated in Prékopa (2013), if the random vector $\tilde{\boldsymbol{\xi}}$ follows a log-concave probability distribution and $g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is quasi-convex, the feasible region defined in CCP (1) is convex. More convexity results can be found in Henrion (2007), Henrion and Strugarek (2008), Lagoa et al. (2005), Henrion and Strugarek (2011). However, it may still be hard to evaluate the probabilistic constraint in CCP (1) precisely even if it is convex. The second method is to consider approximations of chance constraints using the Monte Carlo approach, e.g., sampling average approximation (SAA) proposed by Luedtke and Ahmed (2008). The advantage of SAA is to approximate a chance constraint with the one under finite support with arbitrary accuracy, and the latter can be recast as a mixed-integer program (Ruszczyński 2002). The third method is to propose convex inner approximations of the nonconvex chance constraint (see, e.g., Nemirovski and Shapiro 2006, Calafiore and Campi 2006, Nemirovski and Shapiro 2007). The best-known convex approximation is to replace the chance constraint in CCP (1) with the conditional value-at-risk (CVaR) approximation proposed by Nemirovski and Shapiro (2007). The CVaR approximation usually returns a feasible yet sub-optimal solution. Other nonlinear programming approaches have been developed recently, such as difference-of-convex functions approximation (Hong et al. 2011), a smooth sampling-based approximation (Pena-Ordieres et al. 2020). These approaches often find stationary points of a CCP and thus are not known whether they can be more effective than CVaR approximation or not. In 2017, **A**hmed, **L**uedtke, **SO**ng, and **X**ie (Ahmed et al. 2017) proposed a heuristic scheme, called "ALSO-X" in this paper, which could effectively solve all of their testing instances within the 4% optimality gap. Albeit numerically promising, its theoretical performances are not clear. This paper fills this gap. One main result in this paper is that ALSO-X outperforms the CVaR approximation.

When the distributional information is limited, as a better alternative to the conventional CCPs, distributionally robust chance constrained programs (DRCCPs) have attracted much attention (see, e.g., Hanasusanto et al. 2015, 2017, Xie and Ahmed 2018a, Zymler et al. 2013, Chen et al. 2018, Xie 2019), where the latter is shown to be effective for decision-making under uncertainty without fully knowing the probability distribution. Interested readers are referred to the work (Rahimian and Mehrotra 2019) for a comprehensive review. The particular ambiguity set we focus on in this paper is type $\infty-$Wasserstein ambiguity set.

1.2. Summary of Contributions

In this paper, we study and generalize ALSO-X for solving a CCP and its distributionally robust counterpart (i.e., DRCCP). Our main contributions are summarized below.

(i) We show that when the uncertain constraints are convex, ALSO-X always outperforms CVaR, the well-known best convex approximation, and provide sufficient conditions under which ALSO-X can return an optimal solution to CCP (1).

(ii) We derive an equivalent bilinear programming formulation of CCP (1), which inspires us to improve ALSO-X with a convergent Alternating Minimization (AM) method, termed "ALSO-X+." We show that the solution from the AM method is at least as good as that from the difference-of-convex (DC) approach.

(iii) We extend ALSO-X and CVaR approximation to solve DRCCPs under $\infty-$Wasserstein ambiguity set, termed "the worst-case ALSO-X," and "the worst-case CVaR approximation," respectively. We show that under $\infty-$Wasserstein ambiguity set, the worst-case ALSO-X outperforms the worst-case CVaR approximation.

The roadmap of contributions of our paper is shown in Figure 1.

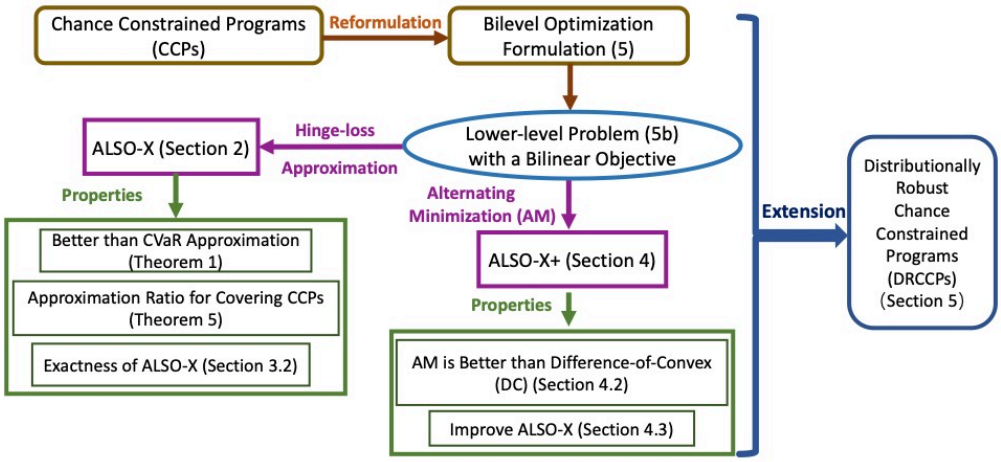

Figure 1 A Roadmap of the Main Results in This Paper.

Organization. The remainder of the paper is organized as follows. Section 2 details the properties of ALSO-X. Section 3 describes the strengths of ALSO-X. Section 4 provides the formulation and properties of ALSO-X+. Section 5 extends and studies ALSO-X and ALSO-X+ to solve DRCCPs under $\infty-$Wasserstein ambiguity set. Section 6 numerically illustrates the proposed methods. Section 7 concludes the paper.

Notation. The following notation is used throughout the paper. We use bold-letters (e.g., $\boldsymbol{x}, \boldsymbol{A}$) to denote vectors and matrices and use corresponding non-bold letters to denote their components. Given a vector or matrix $\boldsymbol{x}$, its zero norm $\|\boldsymbol{x}\|_0$ denotes the number of its nonzero elements. We let $\|\cdot\|_*$ denote the dual norm of a general norm $\|\cdot\|$. We let $\boldsymbol{e}$ be the vector or matrix of all ones, and let $\boldsymbol{e}_i$ be the ith standard basis vector. Given an integer n, we let $[n] := \{1, 2, \ldots, n\}$, and use

$\mathbb{R}^n_+ := \{\boldsymbol{x} \in \mathbb{R}^n : x_i \geq 0, \forall i \in [n]\}$. Given a real number t, we let $(t)_+ := \max\{t, 0\}$. Given a finite set I, we let $|I|$ denote its cardinality. We let $\tilde{\boldsymbol{\xi}}$ denote a random vector and denote its realizations by $\boldsymbol{\xi}$. Given a vector $\boldsymbol{x} \in \mathbb{R}^n$, let $\mathrm{supp}(\boldsymbol{x})$ be its support, i.e., $\mathrm{supp}(\boldsymbol{x}) := \{i \in [n] : x_i \neq 0\}$. Given a probability distribution $\mathbb{P}$ on Ξ, we use $\mathbb{P}\{A\}$ to denote $\mathbb{P}\{\boldsymbol{\xi} : \text{condition } A(\boldsymbol{\xi}) \text{ holds}\}$ when $A(\boldsymbol{\xi})$ is a condition on $\boldsymbol{\xi}$, and to denote $\mathbb{P}\{\boldsymbol{\xi} \colon \boldsymbol{\xi} \in A\}$ when $A \subseteq \Xi$ is $\mathbb{P}-$measurable. We follow the convention of CCP literature (see, e.g., Ruszczyński 2002, Nemirovski and Shapiro 2007, Luedtke and Ahmed 2008) for the definition of an indicator function, that is, given a set R, its normal cone at point $\boldsymbol{x} \in R$ is denoted by $\mathcal{N}_R(\boldsymbol{x}) := \{\boldsymbol{h} : \boldsymbol{h}^\top(\widehat{\boldsymbol{x}} - \boldsymbol{x}) \leq 0, \forall \widehat{\boldsymbol{x}} \in R\}$, and $\emptyset$ if $\boldsymbol{x} \notin R$; and the indicator function $\mathbb{I}(\boldsymbol{x} \in R) = 1$ if $\boldsymbol{x} \in R$, and 0, otherwise. We use $\lfloor x \rfloor$ to denote the largest integer y satisfying $y \leq x$, for any $x \in \mathbb{R}$. We use the phrase "Better Than" to indicate "at least as good as." Additional notation will be introduced as needed.

2. Developments and Properties of ALSO-X

In this section, we present equivalent formulations of CCP (1), derive its hinge-loss approximation, show its connection to the ALSO-X, and we also derive two special cases of ALSO-X.

2.1. Equivalent Formulations

The fact that $\mathbb{P}\{\tilde{\boldsymbol{\xi}} \colon g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0\} = \mathbb{E}_{\mathbb{P}}[\mathbb{I}(g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0)]$ inspires us to introduce a binary functional variable $z(\cdot) : \Xi \to \mathbb{R}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ to represent the indicator function $\mathbb{I}(\cdot)$. Thus, we have the following equivalent formulation of CCP (1),

$$v^* = \min_{\boldsymbol{x} \in \mathcal{X}, z(\cdot)} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \mathbb{I}(g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0) \geq z(\tilde{\boldsymbol{\xi}}), \mathbb{E}[z(\tilde{\boldsymbol{\xi}})] \geq 1 - \varepsilon, z(\tilde{\boldsymbol{\xi}}) \in \{0, 1\} \right\}, \tag{2}$$

where for the sake of simplicity, we suppose that all the random constraints are satisfied almost surely throughout the paper. Next, we observe that CCP (2) can be equivalently reformulated as a bilinear constrained program by introducing another auxiliary nonnegative functional variable $s(\cdot) : \Xi \to \mathbb{R}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ to denote uncertain constraint violations and subsequently replacing random constraint $\mathbb{I}(g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0) \geq z(\tilde{\boldsymbol{\xi}})$ by $\mathbb{E}[z(\tilde{\boldsymbol{\xi}}) s(\tilde{\boldsymbol{\xi}})] = 0$.

Proposition 1 *The CCP* (1) *can be viewed as the following equivalent form*

$$v^* = \min_{\boldsymbol{x} \in \mathcal{X}, z(\cdot), s(\cdot)} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq s(\tilde{\boldsymbol{\xi}}), \mathbb{E}[z(\tilde{\boldsymbol{\xi}})] \geq 1 - \varepsilon, z(\tilde{\boldsymbol{\xi}}) \in [0, 1], \mathbb{E}[z(\tilde{\boldsymbol{\xi}}) s(\tilde{\boldsymbol{\xi}})] = 0, s(\tilde{\boldsymbol{\xi}}) \geq 0 \right\}. \tag{3}$$

Proof. See Appendix A.1. □

We remark that in the bilinear formulation (3), the functional variable $z(\cdot)$ can be either binary or continuous, and this property is useful for deriving ALSO-X+ in Section 4.

Replacing the objective function with an auxiliary variable t, we can rewrite CCP (3) as

$$v^* = \min_{\substack{\boldsymbol{x}\in\mathcal{X},t,\\ z(\cdot),s(\cdot)}} \left\{ t\colon g(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) \le s(\tilde{\boldsymbol{\xi}}), \mathbb{E}[z(\tilde{\boldsymbol{\xi}})] \ge 1-\varepsilon, \mathbb{E}[z(\tilde{\boldsymbol{\xi}})s(\tilde{\boldsymbol{\xi}})] = 0, z(\tilde{\boldsymbol{\xi}}) \in [0,1], s(\tilde{\boldsymbol{\xi}}) \ge 0, \boldsymbol{c}^\top \boldsymbol{x} \le t \right\}. \quad (4)$$

Formulation (4) implies that in a CCP, one can search the smallest possible t such that the constraint system remains feasible. Thus, this motivates us to convert CCP (4) into an equivalent simple bilevel optimization problem.

Proposition 2 *CCP* (4) *is equivalent to*

$$v^* = \min_{t}\ t, \quad (5a)$$

$$s.t.\ (\boldsymbol{x}^*, s^*(\cdot), z^*(\cdot)) \in \arg\min_{\substack{\boldsymbol{x}\in\mathcal{X},\\ z(\cdot)\in[0,1],\\ s(\cdot)\ge 0}} \left\{ \mathbb{E}[z(\tilde{\boldsymbol{\xi}})s(\tilde{\boldsymbol{\xi}})]\colon g(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) \le s(\tilde{\boldsymbol{\xi}}), \mathbb{E}[z(\tilde{\boldsymbol{\xi}})] \ge 1-\varepsilon, \boldsymbol{c}^\top \boldsymbol{x} \le t \right\}, \quad (5b)$$

$$\mathbb{P}\left\{ \tilde{\boldsymbol{\xi}}\colon g(\boldsymbol{x}^*,\tilde{\boldsymbol{\xi}}) \le 0 \right\} \ge 1-\varepsilon. \quad (5c)$$

Proof. See Appendix A.2. □

In the bilevel optimization Formulation (5), the problem defined in (5a) and (5c) is known as an upper-level (or leader's) problem with decision variable t, and the one appearing in (5b) can be regarded as the lower-level (or follower's) problem with decision variables $\boldsymbol{x}$, $s(\cdot)$, and $z(\cdot)$. Given the value of upper-level decision t, we can solve the lower-level problem (5b) and then check whether the solution satisfies condition (5c) or not — if the answer is YES, we can reduce the value of t; otherwise, we have to increase t. Besides, the optimal value of the lower-level problem (5b) is monotone nonincreasing with respect to t. Therefore, if the lower-level problem (5b) were easy to solve, it would be trivial to find the optimal value of CCP (5). For example, the binary search can be used to find the optimal upper-level decision t^*. However, for a given t, the lower-level problem (5b) is a two-stage bilinear program, which is known to be challenging to solve. In fact, solving the original CCP (1) is polynomial-time reducible to problem (5b) since the binary search can take a polynomial number of iterations in the size of the problem. Note that solving a CCP is NP-hard as shown in Luedtke et al. (2010), so is the lower-level problem (5b).

2.2. What is ALSO-X?

As discussed in the previous subsection, the lower-level problem (5b) can be challenging to solve. Thus, in this subsection, we study its hinge-loss approximation, which is much easier to handle in many cases. Particularly, instead of solving a difficult bilinear program, we adopt a simple strategy by letting the functional variable $z(\boldsymbol{\xi}) = 1$ in the lower-level problem (5b), i.e., the following form

$$v^A(t) := \min_{\boldsymbol{x}\in\mathcal{X}, s(\cdot)\ge 0} \left\{ \mathbb{E}[s(\tilde{\boldsymbol{\xi}})]\colon g(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) \le s(\tilde{\boldsymbol{\xi}}), \boldsymbol{c}^\top \boldsymbol{x} \le t \right\}.$$

Next, projecting out continuous variable $s(\cdot)$, we arrive at the following hinge-loss approximation

$$v^A(t) = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \mathbb{E}[g(\boldsymbol{x},\tilde{\boldsymbol{\xi}})_+] \colon \boldsymbol{c}^\top \boldsymbol{x} \le t \right\}. \tag{6}$$

The objective function in (6), termed "the hinge-loss function," can be viewed as an expectation of the nonnegative part of the random function $g(\boldsymbol{x},\tilde{\boldsymbol{\xi}})$, which has been widely applied in machine learning methods such as SVM (see, e.g., Suykens and Vandewalle 1999) and LASSO (see, e.g., Tibshirani 1996). The goal of the hinge-loss approximation (6) is to minimize the expectation of infeasibilities given the upper-level problem's decision t.

The proposed ALSO-X is to replace the lower-level problem (5b) by the hinge-loss approximation (6), which admits the following form

$$v^A = \min_{t} \quad t, \tag{7a}$$

$$\text{s.t.} \quad (\boldsymbol{x}^*, s^*(\cdot)) \in \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X}, s(\cdot)\ge 0} \left\{ \mathbb{E}[s(\tilde{\boldsymbol{\xi}})] \colon g(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) \le s(\tilde{\boldsymbol{\xi}}), \boldsymbol{c}^\top \boldsymbol{x} \le t \right\}, \tag{7b}$$

$$\mathbb{P}\left\{ \tilde{\boldsymbol{\xi}} : s^*(\tilde{\boldsymbol{\xi}}) = 0 \right\} \ge 1-\varepsilon. \tag{7c}$$

Letting $(\boldsymbol{x}^*, s^*(\cdot))$ denote an optimal solution of the hinge-loss approximation (7b), if its optimal solution $\boldsymbol{x}^*$ is feasible to CCP (1), then we have $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : g(\boldsymbol{x}^*,\tilde{\boldsymbol{\xi}}) \le 0\} \ge 1-\varepsilon$, which is equivalent to $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : s^*(\tilde{\boldsymbol{\xi}}) = 0\} \ge 1-\varepsilon$. Increasing t in (7b) would drive down the optimal $s^*(\cdot)$ and hence make (7c) more likely to be satisfied. This ensures that

Proposition 3 *The proposed* ALSO-X (7) *is a convex approximation of CCP* (1)*, i.e.,* $v^A \ge v^*$.

In general, it is difficult to quantify the difference between v^A and v^*, and similar to many conservative approximation methods, under some extreme cases, ALSO-X (7) may be infeasible (please see Example 11 in Appendix F). However, in the next section, we show that for the finite-support covering CCPs, v^A is within a factor (greater than one) of v^* and this factor is tight.

Correspondingly, we develop the ALSO-X Algorithm 1, which generalizes the heuristic algorithm in section 6.1 (Ahmed et al. 2017). Particularly, for a given value t of the upper-level problem, we solve the hinge-loss approximation (7b) with an optimal solution $(\boldsymbol{x}^*, s^*(\cdot))$ and check whether $\boldsymbol{x}^*$ is feasible to CCP (1) or not, i.e., check if $\mathbb{P}\{\tilde{\boldsymbol{\xi}} \colon s^*(\tilde{\boldsymbol{\xi}}) = 0\} \ge 1-\varepsilon$ or not. If the answer is YES, we reduce the value of t. Otherwise, increase it. In the implementation, we search the optimal t by using the binary search method with a proper stopping tolerance δ_1 (e.g., we choose $\delta_1 = 10^{-2}$ in numerical study), which is detailed in Algorithm 1.

We make the following remarks about ALSO-X Algorithm 1.

(i) At Step 4 in Algorithm 1, we let $t_L = t$ if the solution $(\boldsymbol{x}^*, s^*(\cdot))$ is infeasible to the CCP (1), i.e., $\mathbb{P}\{\tilde{\boldsymbol{\xi}} \colon s^*(\tilde{\boldsymbol{\xi}}) > 0\} > \varepsilon$; otherwise, we decrease the current upper bound of t by letting $t_U = t$;

Algorithm 1 The Proposed ALSO-X Algorithm

Input: Let δ_1 denote the stopping tolerance parameter, t_L and t_U be the known lower and upper bounds of the optimal value of CCP (1), respectively
while $t_U - t_L > \delta_1$ **do**
    Let $t = (t_L + t_U)/2$ and $(\boldsymbol{x}^*, s^*(\cdot))$ be an optimal solution of the hinge-loss approximation (7b)
    Let $t_L = t$ if $\mathbb{P}\{\tilde{\boldsymbol{\xi}}\colon s^*(\tilde{\boldsymbol{\xi}}) > 0\} > \varepsilon$; otherwise, $t_U = t$
end while
Output: A feasible solution $\boldsymbol{x}^*$ and its objective value $\bar{v}^A$ to CCP (1)

(ii) For the linear CCPs, i.e., $g_i(\boldsymbol{x}, \boldsymbol{\xi}) = \boldsymbol{\xi}^\top \boldsymbol{a}_i(\boldsymbol{x}) - b_i(\boldsymbol{x})$ with affine functions $\boldsymbol{a}_i(\boldsymbol{x}), b_i(\boldsymbol{x})$ and set $\mathcal{X}$ is a polyhedron, we can use parametric linear programming techniques (see, e.g., Adler and Monteiro 1992) to decrease the number of bisection needed. That is, at each iteration, if the current solution is not feasible to CCP (1), then we update t to be the upper bound of its allowable range; otherwise, we let t be equal to the lower bound of the allowable range. Then, we continue the binary search procedure;

(iii) Note that if the probability distribution $\mathbb{P}$ is finite-support or the probability distribution is elliptical and CCP (1) solely involves a single linear chance constraint, then the hinge-loss approximation (7b) can be efficiently solvable under mild conditions (see Corollary 3 and Corollary 4 in Appendix E). Otherwise, one can first use the sampling average approximation results from Luedtke and Ahmed (2008) to tightly approximate a CCP to the one with finite-support and then solve the ALSO-X (7);

(iv) For nonlinear chance constraints, we can use subgradient descent algorithm to solve the hinge-loss approximation (6), which is detailed in the next subsection;

(v) At each iteration, we can warm-start the process with the solution found in the previous iteration;

(vi) High-quality upper and lower bounds can help reduce the number of iterations needed. For instance, we can use the quantile bound proposed in Ahmed et al. (2017), Song et al. (2014) as a promising lower bound t_L, and use the objective value from CVaR approximation or other heuristics as a potential upper bound t_U; and

(vii) As long as the encoding length of t_U, t_L, and δ_1 are polynomial in the input size of CCP (1), the number of bisection iterations needed in ALSO-X Algorithm 1 is proportional to $\log((t_U - t_L)/\delta_1)$, i.e., polynomial in the input size of CCP (1) as well.

2.3. Subgradient Descent (SD) Algorithm

Note that the hinge-loss approximation (6) is a convex minimization problem and can be efficiently solved by the Subgradient Descent (SD) algorithm if the underlying distribution is finite-support. For ease of notation, we first define the feasibility set of $\boldsymbol{x}$ as $\mathcal{S} := \{\boldsymbol{x}\colon \boldsymbol{x} \in \mathcal{X} \cap \boldsymbol{c}^\top \boldsymbol{x} \le t\}$. The SD method proceeds as follows: (i) Given a solution $\boldsymbol{x} \in \mathcal{S}$, we first find its subgradient; and (ii) then project the difference of this solution and scaled subgradient descent direction into set $\mathcal{S}$, which generates a new solution. We continue this process until invoking a stopping criterion. The detailed SD method for solving the hinge-loss approximation (6) can be found in Algorithm 2.

Algorithm 2 Subgradient Descent (SD) Algorithm to Solve the Hinge-loss Approximation (6)

1: Find an initial feasible solution $\boldsymbol{x}_0$ such that $\boldsymbol{x}_0 \in \mathcal{S} := \{\boldsymbol{x}\colon \boldsymbol{x} \in \mathcal{X} \cap \boldsymbol{c}^\top \boldsymbol{x} \le t\}$ and let $k = 0$
2: **do**
3: At iteration k, compute subgradient $\widehat{h}_{\boldsymbol{x}_k}$ for $\boldsymbol{x}_k$, i.e., $\widehat{h}_{\boldsymbol{x}_k} = \partial_{\boldsymbol{x}_k} \mathbb{E}[g(\boldsymbol{x}_k, \tilde{\boldsymbol{\xi}})_+]$
4: Update $\boldsymbol{x}_{k+1}$ by $\Pi_{\mathcal{S}}(\boldsymbol{x}_k - \gamma_k \widehat{h}_{\boldsymbol{x}_k})$, where γ_k is the step size
5: $k = k + 1$
6: **while** Invoking a stopping criterion

We make the following remarks about SD Algorithm 2.

(i) According to Assumption A1 and theorem 1 in Rockafellar and Wets (1982), we can interchange the subdifferential operator and expectation when updating $\boldsymbol{x}_k$ at kth iteration. That is, $\boldsymbol{x}_{k+1} = \Pi_{\mathcal{S}}(\boldsymbol{x}_k - \gamma_k \mathbb{E}[\partial_{\boldsymbol{x}_k} g(\boldsymbol{x}_k, \tilde{\boldsymbol{\xi}})_+])$;

(ii) The proposed Algorithm 2 works well when the underlying probability is finite-support and the random function $g_i(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is convex and nonlinear. For the continuous probability distribution $\mathbb{P}$, one can approximate it with the one of finite support, according to the sampling average approximation results in Luedtke and Ahmed (2008), or one can use the stochastic subgradient descent method (Nemirovski et al. 2009);

(iii) At Step 4 in Algorithm 2, the projection onto the feasible set $\mathcal{S}$ can be solved efficiently by adopting Dykstra's projection algorithm (Boyle and Dykstra 1986), i.e., projecting onto set $\mathcal{X}$ and the set induced by the knapsack constraint (i.e., set $\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{c}^\top \boldsymbol{x} \le t\}$) alternatively (Tavakoli 2016); and

(iv) In the numerical study, we select the step size at iteration k as $\gamma_k = 1/(k+1)$, and its corresponding convergence rate is $\mathcal{O}(1/\log(T))$. Other step size choices and convergence rate results can be found in chapter 3 (Nesterov 2003).

2.4. Special Cases

In this subsection, we discuss two special cases under which the hinge-loss approximation (6) can be computed efficiently: (i) under discrete support; or (ii) if $I=1$, and function $g(\boldsymbol{x},\boldsymbol{\xi})$ is biaffine in $\boldsymbol{x}$ and $\boldsymbol{\xi}$ and the random vector $\tilde{\boldsymbol{\xi}}$ follows an elliptical distribution.

Special Case 1: Discrete Support

If the underlying probability distribution is finite-support with N equiprobable scenarios, i.e., the random vector $\tilde{\boldsymbol{\xi}}$ has a finite support $\Xi=\{\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^N\}$ with $\mathbb{P}\{\tilde{\boldsymbol{\xi}}=\boldsymbol{\xi}^i\}=1/N$ for all $i\in[N]$, then CCP (1) reduces to

$$v^*=\min_{\boldsymbol{x}\in\mathcal{X}}\left\{\boldsymbol{c}^\top\boldsymbol{x}\colon \sum_{i\in[N]}\mathbb{I}(g(\boldsymbol{x},\boldsymbol{\xi}^i)\le 0)\ge N-\lfloor N\varepsilon\rfloor\right\},$$

and ALSO-X (7) admits the following form

$$\begin{aligned} v^A=\min_{t}\quad & t,\\ \text{s.t.}\quad & (\boldsymbol{x}^*,\boldsymbol{s}^*)\in\operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{s}\ge\boldsymbol{0}}\left\{\frac{1}{N}\sum_{i\in[N]}s_i\colon \boldsymbol{c}^\top\boldsymbol{x}\le t, g(\boldsymbol{x},\boldsymbol{\xi}^i)\le s_i,\forall i\in[N]\right\},\\ & \sum_{i\in[N]}\mathbb{I}(s_i^*=0)\ge N-\lfloor N\varepsilon\rfloor. \end{aligned} \tag{8}$$

Note that for this special case, the condition that $\mathbb{P}\{\tilde{\boldsymbol{\xi}}\colon s^*(\tilde{\boldsymbol{\xi}})=0\}\ge 1-\varepsilon$ in ALSO-X (7c) reduces to that $\sum_{i\in[N]}\mathbb{I}(s_i^*=0)\ge N-\lfloor N\varepsilon\rfloor$, where the left-hand side of the inequality is equal to the support size of $\boldsymbol{s}^*$ (i.e., $|\mathrm{supp}(\boldsymbol{s}^*)|$). Thus, if the support size of current solution $\boldsymbol{s}^*$, or equivalently, the number of violations is strictly larger than $\lfloor N\varepsilon\rfloor$, then the current solution $\boldsymbol{x}^*$ is not feasible to CCP (1). The ALSO-X (8) can be generalized to the case when the probability mass is not uniform (see Section E.1 in Appendix E for the detailed formulation).

Special Case 2: Elliptical Distributions

Elliptical distributions have been widely used in risk management (see, e.g., Landsman and Valdez 2003, Jaworski et al. 2010, Embrechts et al. 2002, Kamdem 2005). For instance, Kamdem 2005 used elliptical distributions to model portfolio risk factors. An elliptical distribution $\mathbb{P}_{\mathrm{E}}(\boldsymbol{\mu},\boldsymbol{\Sigma},\widehat{g})$ is described by three parameters, a location parameter $\boldsymbol{\mu}$, a positive definite matrix $\boldsymbol{\Sigma}$, and a generating function $\widehat{g}$. The name of an elliptical distribution is based on the fact that the contours of its density are ellipsoids in $\boldsymbol{x}\in\mathbb{R}^n$, and therefore, its probability density function $\widehat{f}$ is defined as

$$\widehat{f}(\boldsymbol{x})=k\cdot\widehat{g}\left(\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

with a positive normalization scalar k.

The probability density function of the standard univariate elliptical distribution $\mathbb{P}_{\mathrm{E}}(0,1,\widehat{g})$ is

$\varphi(z) = k\widehat{g}(z^2/2)$, and the corresponding cumulative distribution function is $\Phi(a) = \int_{-\infty}^{a} k\widehat{g}(z^2/2)dz$. For the single linear CCP (1), i.e., $I = 1$ and $g(\boldsymbol{x}, \boldsymbol{\xi}) = \boldsymbol{\xi}^\top \boldsymbol{a}_1(\boldsymbol{x}) - b_1(\boldsymbol{x})$ with affine functions $\boldsymbol{a}_1(\boldsymbol{x})$, $b_1(\boldsymbol{x})$, if the random parameters $\tilde{\boldsymbol{\xi}}$ follow a joint elliptical distribution with $\tilde{\boldsymbol{\xi}} \sim \mathbb{P}_{\mathrm{E}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{g})$, the objective function in the hinge-loss approximation (6) can be much simplified. In this special case, CCP (1) reduces to the following conic program (see, e.g., problem 2.1 in Kataoka 1963 or theorem 3 in Prékopa 1974)

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon b_1(\boldsymbol{x}) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x}) \geq \Phi^{-1}(1-\varepsilon)\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x})} \right\}. \tag{9}$$

This notable simplification (9) is useful to show the exactness of ALSO-X (7). The following proposition shows an equivalent reformulation of ALSO-X (7) .

Proposition 4 *For any elliptical distribution* $\mathbb{P}_{\mathrm{E}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{g})$, ALSO-X (7) *corresponding to the single linear CCP admits the following form*

$$v^A = \min_{t} \quad t, \tag{10a}$$

$$\begin{aligned} s.t. \quad (\boldsymbol{x}^*, \alpha^*) \in \arg\min_{\boldsymbol{x}\in\mathcal{X},\alpha} \Bigg\{ & \sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x})} \left(\overline{G}(\alpha^2/2) - \alpha + \alpha\Phi(\alpha) \right) : \\ & \boldsymbol{c}^\top \boldsymbol{x} \leq t, \frac{b_1(\boldsymbol{x}) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x})}{\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x})}} = \alpha \Bigg\}, \end{aligned} \tag{10b}$$

$$b_1(\boldsymbol{x}^*) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x}^*) \geq \Phi^{-1}(1-\varepsilon)\sqrt{\boldsymbol{a}_1(\boldsymbol{x}^*)^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x}^*)}, \tag{10c}$$

where $\overline{G}(\tau) = G(\infty) - G(\tau)$ *and* $G(\tau) = k\int_0^\tau \widehat{g}(z)dz$. *By default, we let* $\frac{0}{0} = 0$ *and* $\frac{c}{0} = \mathrm{sign}(c)\infty$ *if* $c \neq 0$.

Proof. See Appendix A.3. □

Note that one might need to project out variable α to ensure the convexity of the objective function in the lower-level problem (10b). The purpose of introducing variable α is to simplify the formula and is convenient to prove the monotonicity of the objective function, which is elaborated in the next section.

The following result shows under a Gaussian distribution, the objective function of the hinge-loss approximation (6) can be further simplified according to Proposition 4.

Corollary 1 *When* $\tilde{\boldsymbol{\xi}}$ *follows Gaussian distribution (i.e., a special elliptical distribution with* $\widehat{g}(\mu) = e^{-\mu}$ *and* $k = (\sqrt{2\pi})^{-1}$*), the hinge-loss approximation* (10b) *corresponding to the single linear CCP reduces to*

$$v^A(t) = \min_{\boldsymbol{x}\in\mathcal{X},\alpha} \left\{ \sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x})} \left(\varphi(\alpha) - \alpha + \alpha\Phi(\alpha) \right) : \boldsymbol{c}^\top \boldsymbol{x} \leq t, \frac{b_1(\boldsymbol{x}) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x})}{\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x})}} = \alpha \right\}. \tag{11}$$

Finally, we re-stress that Proposition 4 and Corollary 1 are useful to derive the exactness of ALSO-X.

3. Strengths of ALSO-X

In this section, we present three strengths of ALSO-X. First, we demonstrate ALSO-X (7) always outperforms CVaR approximation under Assumptions A1 and A2. Next, we show sufficient conditions under which ALSO-X (7) returns an optimal solution to CCP (1). Finally, we provide a provable performance guarantee when applying the proposed ALSO-X (7) to solve the covering CCPs.

3.1. A Comparison between ALSO-X and CVaR approximation

Given a random variable $\tilde{\boldsymbol{X}}$, let $\mathbb{P}$ and $F_{\tilde{\boldsymbol{X}}}(\cdot)$ be its probability distribution and cumulative distribution function, respectively. For a given risk level ε, $(1-\varepsilon)-$Value at risk (VaR) of $\tilde{\boldsymbol{X}}$ is

$$\mathrm{VaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}) := \min_s \{s : F_{\tilde{\boldsymbol{X}}}(s) \geq 1-\varepsilon\},$$

and the corresponding conditional value-at-risk (CVaR) is defined as

$$\mathrm{CVaR}_{1-\varepsilon}(\tilde{\boldsymbol{X}}) := \min_{\beta} \left\{ \beta + \frac{1}{\varepsilon}\mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{X}} - \beta]_+ \right\}.$$

According to Nemirovski and Shapiro (2007), the CVaR approximation of CCP (1) can be written as

$$v^{\mathrm{CVaR}} = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \min_{\beta\leq 0} \left\{ \beta + \frac{1}{\varepsilon}\mathbb{E}\{g(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) - \beta\}_+ \right\} \leq 0 \right\}. \tag{12a}$$

Letting $s(\boldsymbol{\xi}) := \max\{g(\boldsymbol{x},\boldsymbol{\xi}), \beta\}$ and linearizing it, the CVaR approximation (12a) is equivalent to

$$v^{\mathrm{CVaR}} = \min_{\boldsymbol{x}\in\mathcal{X},\beta\leq 0, s(\cdot)} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon g(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) \leq s(\tilde{\boldsymbol{\xi}}), \mathbb{E}[s(\tilde{\boldsymbol{\xi}})] - (1-\varepsilon)\beta \leq 0, s(\tilde{\boldsymbol{\xi}}) \geq \beta \right\}. \tag{12b}$$

Next, augmenting the objective function by adding an auxiliary variable t, the CVaR approximation (12b) can be further formulated as

$$v^{\mathrm{CVaR}} = \min_{\boldsymbol{x}\in\mathcal{X},\beta\leq 0, s(\cdot), t} \left\{ t \colon g(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) \leq s(\tilde{\boldsymbol{\xi}}), \mathbb{E}[s(\tilde{\boldsymbol{\xi}})] - (1-\varepsilon)\beta \leq 0, s(\tilde{\boldsymbol{\xi}}) \geq \beta, \boldsymbol{c}^\top \boldsymbol{x} \leq t \right\}. \tag{12c}$$

Similar to ALSO-X (7), in the CVaR approximation (12c), one can search the smallest possible t such that the constraint $\mathbb{E}[s(\tilde{\boldsymbol{\xi}})] - (1-\varepsilon)\beta$ remains feasible. Thus, the CVaR approximation (12c) can be rewritten as the following simple bilevel program

$$v^{\mathrm{CVaR}} = \min_t \ t, \tag{13a}$$

$$\text{s.t.} \quad (\boldsymbol{x}^*, s^*(\cdot), \beta^*) \in \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{X},\beta\leq 0, s(\cdot)} \left\{ \mathbb{E}[s(\tilde{\boldsymbol{\xi}})] - (1-\varepsilon)\beta \colon g(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) \leq s(\tilde{\boldsymbol{\xi}}), s(\tilde{\boldsymbol{\xi}}) \geq \beta, \boldsymbol{c}^\top \boldsymbol{x} \leq t \right\}, \tag{13b}$$

$$\mathbb{E}[s^*(\tilde{\boldsymbol{\xi}})] - (1-\varepsilon)\beta^* \leq 0. \tag{13c}$$

Above, for any given t, let $(\boldsymbol{x}^*, s^*(\cdot), \beta^*)$ be an optimal solution of the lower-level problem (13b) with an optimal objective value $v^{\mathrm{CVaR}}(t)$, where

$$v^{\mathrm{CVaR}}(t) = \min_{\boldsymbol{x}\in\mathcal{X}, \beta\le 0, s(\cdot)} \left\{ \mathbb{E}[s(\tilde{\boldsymbol{\xi}})] - (1-\varepsilon)\beta \colon g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \le s(\tilde{\boldsymbol{\xi}}), s(\tilde{\boldsymbol{\xi}}) \ge \beta, \boldsymbol{c}^\top \boldsymbol{x} \le t \right\}.$$

If $v^{\mathrm{CVaR}}(t) > 0$, then constraint (13c) would be violated and $(\boldsymbol{x}^*, s^*(\cdot), \beta^*, t)$ would be infeasible to the CVaR approximation (12c). As a result, we must have $t < v^{\mathrm{CVaR}}$. Otherwise, we would have $t \ge v^{\mathrm{CVaR}}$.

Notice that letting the variable $\beta = 0$ in the lower-level problem (13b) recovers the hinge-loss approximation (7b). This observation motivates us to compare CVaR approximation (13) and ALSO-X (7); namely, for a given t, if an optimal solution of the lower-level problem (7b) violates the chance constraint (7c), so does the CVaR approximation. In fact, we can prove that the optimal values of both lower-level problems coincide under this premise. Therefore, the ALSO-X (7) outperforms the CVaR approximation (13), since the feasibility-checking condition (i.e., constraint (13c)) of the upper-level problem in the CVaR approximation (13) is more restrictive.

Theorem 1 *Let v^A, v^{CVaR} denote the optimal value of the* ALSO-X (7) *and the* CVaR *approximation* (13)*, respectively. Then, under Assumptions A1-A2, we must have $v^A \le v^{\mathrm{CVaR}}$.*

Proof. See Appendix A.4. □

Interested readers are referred to Example 7 in Appendix B for an illustration of the correctness of Theorem 1.

We note that according to the stopping criterion in Algorithm 1, the output objective value $\bar{v}^A$ might not be equal to v^A and at most δ_1 larger than v^A, the optimal value of ALSO-X. According to Theorem 1, we must have $v^A \le v^{\mathrm{CVaR}}$. Thus, the output objective value $\bar{v}^A$ from Algorithm 1 is no larger than $v^{\mathrm{CVaR}} + \delta_1$. This result is summarized below.

Corollary 2 *Under Assumptions A1-A2,* ALSO-X *Algorithm 1 returns a feasible solution with $\bar{v}^A \le v^{\mathrm{CVaR}} + \delta_1$, where $\bar{v}^A$ is the output objective value and δ_1 is the chosen stopping tolerance parameter in Algorithm 1.*

It is worth noting that using a proper stopping tolerance parameter, ALSO-X Algorithm 1 in general returns a better solution than the CVaR approximation; however, since the binary search procedure requires solving many similar hinge-loss approximations, it might be slower than the CVaR approximation. Our numerical study shows that for the linear CCP, the computational time of ALSO-X Algorithm 1 is usually longer than the CVaR approximation since the off-the-shelf solvers excel in solving large-scale linear programs, while for the nonlinear CCP, ALSO-X

Algorithm 1 takes a shorter time than the CVaR approximation using the subgradient method. Nevertheless, in both cases, the solution quality of ALSO-X Algorithm 1 is found to be consistently better than that of the CVaR approximation.

Notably, the convexity assumption of set $\mathcal{X}$ in Assumption A2 is of a necessity to the result of Theorem 1. In the non-convex setting, the ALSO-X (7) cannot be guaranteed to be better than the CVaR approximation (13). Particularly, the following two examples show that the ALSO-X can return a better solution than the CVaR approximation and vice versa when set $\mathcal{X}$ is nonconvex.

Example 1 Let us revisit Example 7 in Appendix B with an additional restriction that x is an integer, i.e, set $\mathcal{X}=\mathbb{Z}_+$. In this case, we have $v^*=2$, $v^A=2$, and $v^{\rm CVaR}=3$. Thus, the ALSO-X outperforms the CVaR approximation. ◇

Example 2 Consider a CCP with 4 equiprobable scenarios (i.e., $N=4$, $\mathbb{P}\{\tilde{\boldsymbol{\xi}}=\boldsymbol{\xi}^i\}=1/N$), risk level $\varepsilon=1/2$, set $\mathcal{X}=\{0,1\}$, function $g(\boldsymbol{x},\boldsymbol{\xi})=\xi_1 x-\xi_2$, $\xi_1^1=-49$, $\xi_1^2=\xi_1^3=\xi_1^4=101$, $\xi_2^1=-50$, and $\xi_2^2=\xi_2^3=\xi_2^4=99$. Under this setting, CCP (1) becomes

$$v^*=\min_{x\in\{0,1\}}\left\{-x\colon \mathbb{I}(49x\geq 50)+\mathbb{I}(101x\leq 99)+\mathbb{I}(101x\leq 99)+\mathbb{I}(101x\leq 99)\geq 2\right\}.$$

The CVaR approximation of this CCP is

$$v^{\rm CVaR}=\min_{x\in\{0,1\},\beta\leq 0,\boldsymbol{s}}\left\{-x\colon \begin{array}{c}-49x+50\leq s_1, 101x-99\leq s_2, 101x-99\leq s_3,\\ 101x-99\leq s_4, \frac{1}{4}\sum_{i\in[4]}s_i-\beta/2\leq 0, s_i\geq\beta, \forall i\in[4]\end{array}\right\}.$$

By the simple calculation, the optimal value is $v^*=0$ with the optimal solution $x^*=0$, and $v^{\rm CVaR}=0$ for the CVaR approximation (12b) with the optimal solution $x^*=0,\beta^*=-123.5,s_1^*=50,s_2^*=s_3^*=s_4^*=-99$. However, the ALSO-X (7) of this example is infeasible, which can be formulated as

$$\begin{aligned}&v^A=\min_t\left\{t\colon \sum_{i\in[4]}\mathbb{I}(s_i^*=0)\geq 2,\right.\\ &\left.(x^*,\boldsymbol{s}^*)\in\mathop{\arg\min}_{x\in\{0,1\},\boldsymbol{s}\geq\mathbf{0}}\left\{\frac{1}{4}\sum_{i\in[4]}s_i\colon \begin{array}{l}-49x+50\leq s_1, 101x-99\leq s_2,\\ 101x-99\leq s_3, 101x-99\leq s_4, -x\leq t\end{array}\right\}\right\}.\end{aligned}$$

Particularly, for any $t\geq -1$, the hinge-loss approximation returns a solution with $s_1^*=1$, $s_2^*=s_3^*=s_4^*=2$, $x^*=1$, and the support size of $\boldsymbol{s}^*$ is greater than 2, then we have to increase the objective bound t to the infinity. Therefore, in this example, the CVaR approximation always returns the optimal solution, but the ALSO-X fails to find any feasible solution. ◇

3.2. Exactness of ALSO-X

In this subsection, we show sufficient conditions under which ALSO-X (7) can provide an exact optimal solution to CCP (1). To begin with, the following example shows that ALSO-X (7) may not be able to find the exact solution of CCP (1) even under Assumptions A1-A2. Thus, in this subsection, we explore the conditions under which the ALSO-X (7) returns an exact optimal solution.

Example 3 Consider a CCP with 3 equiprobable scenarios (i.e., $N=3$, $\mathbb{P}\{\tilde{\boldsymbol{\xi}}=\boldsymbol{\xi}^i\}=1/N$), risk level $\varepsilon=1/3$, set $\mathcal{X}=\mathbb{R}^2_+$, function $g(\boldsymbol{x},\boldsymbol{\xi})=-\boldsymbol{\xi}^\top\boldsymbol{x}+1$, and $\boldsymbol{\xi}^1=(2,3)^\top$, $\boldsymbol{\xi}^2=(2,1)^\top$, $\boldsymbol{\xi}^3=(1,2)^\top$. The optimal value of this CCP can be found by solving the following mixed-integer linear program

$$v^*=\min_{\boldsymbol{x}\in\mathbb{R}^2_+,\boldsymbol{z}\in\{0,1\}^3}\left\{x_1+x_2\colon 2x_1+3x_2\geq z_1, 2x_1+x_2\geq z_2, x_1+2x_2\geq z_3, \sum_{i\in[3]}z_i\geq 2\right\},$$

i.e., $v^*=0.5$.

The corresponding ALSO-X (7) is

$$\begin{aligned}&v^A=\min_t\left\{t\colon \sum_{i\in[3]}\mathbb{I}(s_i^*=0)\geq 2,\right.\\&\left.(\boldsymbol{x}^*,\boldsymbol{s}^*)\in\operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^2_+,\boldsymbol{s}\in\mathbb{R}^3_+}\left\{\frac{1}{3}\sum_{i\in[3]}s_i\colon \begin{array}{l}2x_1+3x_2\geq 1-s_1, 2x_1+x_2\geq 1-s_2,\\ x_1+2x_2\geq 1-s_3, x_1+x_2\leq t\end{array}\right\}\right\}.\end{aligned}$$

Simple calculations show that ALSO-X has an optimal $v^A=2/3>v^*$. ◇

Example 3 motivates us to find the special cases of CCP (1) under which ALSO-X (7) can provide an optimal solution.

Special Case I of Exactness: CCPs with Equality Constraint

This special case of CCP (1) consists of a linear uncertain equality constraint. This special case is motivated by the following two distinct applications:

- The first one is to find a feasible subsystem of linear equalities studied by Amaldi and Kann (1995). That is, given a possibly infeasible linear system $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ with matrix $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ and vector $\boldsymbol{b}\in\mathbb{R}^m$, and a positive integer $K\in[m]$, the goal of the problem is to seek a solution $\boldsymbol{x}\in\mathbb{R}^n$ such that $\boldsymbol{x}$ satisfies at least K linear equalities of the system; and
- The second one is to find a sparse solution from the linear system studied by Nemirovski (2001). That is, given a linear system $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ with matrix $\boldsymbol{A}\in\mathbb{R}^{m\times n}$ and vector $\boldsymbol{b}\in\mathbb{R}^m$ and an integer $K\in[n]$, the goal of the problem is to seek a sparse solution $\boldsymbol{x}$ such that the support size of $\boldsymbol{x}$ is no larger than K.

In this special case, we assume set $\mathcal{X}=\{\boldsymbol{x}\in\mathbb{R}^n:\boldsymbol{U}^\top\boldsymbol{x}=\boldsymbol{h}\}$ with matrix $\boldsymbol{U}\in\mathbb{R}^{m\times n}$ and vector $\boldsymbol{h}\in\mathbb{R}^n$, $g_1(\boldsymbol{x},\boldsymbol{\xi})=\boldsymbol{\xi}^\top(\boldsymbol{A}\boldsymbol{x})+a(\boldsymbol{\xi})-\boldsymbol{b}^\top\boldsymbol{x}$, and $g_2(\boldsymbol{x},\boldsymbol{\xi})=-\boldsymbol{\xi}^\top(\boldsymbol{A}\boldsymbol{x})-a(\boldsymbol{\xi})+\boldsymbol{b}^\top\boldsymbol{x}$, where $\boldsymbol{A}\in\mathbb{R}^{m\times n},\boldsymbol{b}\in\mathbb{R}^n$, and $a(\cdot):\Xi\to\mathbb{R}$. Hence, CCP (1) reduces to

$$v^*=\min_{\boldsymbol{x}\in\mathcal{X}}\left\{\boldsymbol{c}^\top\boldsymbol{x}\colon\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon\tilde{\boldsymbol{\xi}}^\top(\boldsymbol{A}\boldsymbol{x})+a(\tilde{\boldsymbol{\xi}})=\boldsymbol{b}^\top\boldsymbol{x}\right\}\geq 1-\varepsilon\right\},\tag{15}$$

while the corresponding ALSO-X (7) can be written as

$$v^A=\min_{t}\quad t,\tag{16a}$$

$$\text{s.t.}\quad(\boldsymbol{x}^*,s^*(\cdot))\in\arg\min_{\boldsymbol{x}\in\mathcal{X},s(\cdot)}\left\{\mathbb{E}\left[|s(\tilde{\boldsymbol{\xi}})|\right]:\tilde{\boldsymbol{\xi}}^\top(\boldsymbol{A}\boldsymbol{x})+a(\tilde{\boldsymbol{\xi}})-\boldsymbol{b}^\top\boldsymbol{x}=s(\tilde{\boldsymbol{\xi}}),\boldsymbol{c}^\top\boldsymbol{x}=t\right\},\tag{16b}$$

$$\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}:s^*(\tilde{\boldsymbol{\xi}})=0\right\}\geq 1-\varepsilon.\tag{16c}$$

Above, we replace the inequality $\boldsymbol{c}^\top\boldsymbol{x}\leq t$ by the equality $\boldsymbol{c}^\top\boldsymbol{x}=t$ since at optimality, the equality must hold. The sufficient condition under which ALSO-X (16) returns an optimal solution of CCP (15) relies on the following property.

Generalized Nullspace Property. *For any $(\boldsymbol{x},s(\cdot))$ such that $\boldsymbol{\xi}^\top(\boldsymbol{A}\boldsymbol{x})-\boldsymbol{b}^\top\boldsymbol{x}-s(\boldsymbol{\xi})=0,\boldsymbol{c}^\top\boldsymbol{x}=0,\boldsymbol{U}^\top\boldsymbol{x}=\boldsymbol{0},s(\boldsymbol{\xi})\neq 0$, then for any $\mathbb{P}-$measurable set $\mathcal{S}\subseteq\Xi$ such that $\mathbb{P}\{\tilde{\boldsymbol{\xi}}\colon\tilde{\boldsymbol{\xi}}\in\mathcal{S}\}\leq\varepsilon$, one must have $\mathbb{E}[|s(\tilde{\boldsymbol{\xi}})|\mathbb{I}(\tilde{\boldsymbol{\xi}}\in\mathcal{S})]<1/2\mathbb{E}[|s(\tilde{\boldsymbol{\xi}})|]$.*

It is worthy of mentioning that this generalized nullspace property extends the notion of the nullspace property for the sparse signal recovery (i.e., property 1.3.4 in Nemirovski 2001), where the latter is useful to characterize the uniqueness of the sparse solution satisfying a finite set of linear equations. We also remark that if the probability distribution $\mathbb{P}$ consists of N equiprobable scenarios, the generalized nullspace property can be simplified as

Generalized Nullspace Property with N Equiprobable Scenarios. *For any $(\boldsymbol{x},\boldsymbol{s})\in\mathbb{R}^n\times\mathbb{R}^N$ such that ${\boldsymbol{\xi}^i}^\top(A\boldsymbol{x})-B\boldsymbol{x}-s_i=0$ for each $i\in[N]$, $\boldsymbol{c}^\top\boldsymbol{x}=0,\boldsymbol{U}^\top\boldsymbol{x}=\boldsymbol{0},\boldsymbol{s}\neq\boldsymbol{0}$, then for any set $\mathcal{S}\subseteq[N]$ with $|\mathcal{S}|\leq\lfloor\varepsilon N\rfloor$, we must have $\sum_{i\in\mathcal{S}}|s_i|<1/2\sum_{i\in[N]}|s_i|$.*

Next, we are ready to establish the exactness of ALSO-X (16) for this special case, given that the generalized nullspace property holds.

Theorem 2 *For Special Case 1, the following results must hold:*

(i) For any feasible pair of $(t,a(\cdot))$ under which the hinge-loss approximation (16b) *has a feasible solution $(\boldsymbol{x},s(\cdot))$ satisfying $\mathbb{P}\{\tilde{\boldsymbol{\xi}}:s(\tilde{\boldsymbol{\xi}})=0\}\geq 1-\varepsilon$, every optimal solution $(\boldsymbol{x}^*,s^*(\cdot))$ to the hinge-loss approximation* (16b) *shares the same $s^*(\cdot)$ and satisfies $\mathbb{P}\{\tilde{\boldsymbol{\xi}}:s^*(\tilde{\boldsymbol{\xi}})=0\}\geq 1-\varepsilon$, if and only if the generalized nullspace property holds; and*

(ii) Suppose that the generalized nullspace property holds. Then the optimal values of CCP (15) *and* ALSO-X (16) *coincide, i.e., $v^A=v^*$. Moreover, every optimal solution $(\boldsymbol{x}^*,s^*(\cdot))$ to the CCP* (15) *shares the same $s^*(\cdot)$.*

Proof. See Appendix A.5. □

The follow example illustrates the correctness of Theorem 2.

Example 4 Consider a CCP with 3 equiprobable scenarios (i.e., $N=3$, $\mathbb{P}\{\tilde{\boldsymbol{\xi}}=\boldsymbol{\xi}^i\}=1/N$), risk level $\varepsilon=1/3$, set $\mathcal{X}=\mathbb{R}^2$, function $g_1(\boldsymbol{x},\boldsymbol{\xi})=-\boldsymbol{\xi}^\top\boldsymbol{x}+1$ and $g_2(\boldsymbol{x},\boldsymbol{\xi})=\boldsymbol{\xi}^\top\boldsymbol{x}-1$, $\boldsymbol{\xi}^1=(2,3)^\top$, $\boldsymbol{\xi}^2=(2,1)^\top$, $\boldsymbol{\xi}^3=(1,2)^\top$. Under this setting, the optimal solution of this CCP can be obtained by solving the following mixed-integer linear program

$$v^*=\min_{\boldsymbol{x}\in\mathbb{R}^2}\{x_1+x_2\colon \mathbb{I}(2x_1+3x_2=1)+\mathbb{I}(2x_1+x_2=1)+\mathbb{I}(x_1+2x_2=1)\geq 2\}$$

with optimal value $v^*=1/2$.

In this example, the generalized nullspace property holds. In fact, any $(\boldsymbol{x},\boldsymbol{s})\in\mathbb{R}^2\times\mathbb{R}^3$ satisfies the following conditions

$$2x_1+3x_2-s_1=0, 2x_1+x_2-s_2=0, x_1+2x_2-s_3=0, x_1+x_2=0, \boldsymbol{s}\neq\mathbf{0},$$

which is equivalent to

$$x_2=-x_1, s_1=-x_1, s_2=x_1, s_3=-x_1, x_1\neq 0.$$

Simple calculations show that for any set $\mathcal{S}\subseteq[3]$ with $|\mathcal{S}|\leq 1$, we must have $\sum_{i\in\mathcal{S}}|s_i|<1/2\sum_{i\in[3]}|s_i|$. Therefore, according to Theorem 2, the optimal value for ALSO-X (7) must be $v^A=v^*=1/2$. Indeed, in this example, the ALSO-X (7) reduces to

$$\begin{aligned}
v^A=\min_t\Bigg\{t\colon\ &\sum_{i\in[3]}\mathbb{I}(s_i^*=0)\geq 2,\\
&(\boldsymbol{x}^*,\boldsymbol{s}^*)\in\operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^2,\boldsymbol{s}\in\mathbb{R}^3}\left\{\frac{1}{3}\sum_{i\in[3]}|s_i|\colon \begin{array}{l}2x_1+3x_2=1+s_1, 2x_1+x_2=1+s_2,\\ x_1+2x_2=1+s_3, x_1+x_2\leq t\end{array}\right\}\Bigg\}.
\end{aligned}$$

We see that if $t\geq 1/2$, the optimal solution of the corresponding hinge-loss approximation is $x_1^*=1/2$, $x_2^*=0, s_1^*=s_2^*=0, s_3^*=-1/2$. This suggests that the optimal value for ALSO-X (7) is $v^A=v^*=1/2$. ◇

Special Case II of Exactness: CCPs with Generalized Set-covering Type of Uncertain Constraints

In this special case, we consider the function $g\colon\mathcal{X}\times\Xi\to\mathbb{R}_-\cup\{M\}$, where $M\in\mathbb{R}_{++}$ is a positive constant. This special case is a generalization of the chance constrained set covering problem (see, e.g., Ahmed and Papageorgiou 2013, Beraldi and Ruszczyński 2002), where $\mathcal{X}\subseteq\{0,1\}^n$, $M=1$, and $g(\boldsymbol{x},\boldsymbol{\xi})=1-\boldsymbol{\xi}^\top\boldsymbol{x}$ with binary support $\boldsymbol{\xi}\in\{0,1\}^n:=\Xi$. It is worthy of noting that (i) this special case might violate Assumption A2 that set $\mathcal{X}$ is convex and (ii) when the probability distribution is finite-support, this special case has been studied in Ahmed et al. (2017) (see proposition 12). We show that ALSO-X (7) still provides an optimal solution when the probability distribution is arbitrary.

Theorem 3 *(A generalization of proposition 12 in Ahmed et al. 2017) Suppose that* $g(\boldsymbol{x},\boldsymbol{\xi})\colon \mathcal{X}\times\Xi\to\mathbb{R}_-\cup\{M\}$*, where* $M\in\mathbb{R}_{++}$ *is a positive constant, the optimal value of* ALSO-X (7) *coincides with that of CCP* (1)*.*

Proof. See Appendix A.6. □

Interested readers are referred to Example 8 in Appendix B for a demonstration of the Special Case 2.

Special Case III of Exactness: A Single Linear CCP under an Elliptical Distribution Let us revisit Special Case 2 in Section 2.4, which considers a single linear CCP under an elliptical distribution. For this special case, we show that under additional assumptions, ALSO-X (10) provides an optimal solution to CCP (1).

Theorem 4 *For the single linear CCP* (9) *under an elliptical distribution, the* ALSO-X (10) *provides an optimal solution to CCP* (9)*, provided that (i)* $\mathcal{X}\subseteq\{\boldsymbol{x}:\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}=C\}$*, where* C *is a positive constant; or (ii)* $\mathcal{X}\subseteq\{\boldsymbol{x}:b_1(\boldsymbol{x})-\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x})=C\}$*, where* C *is an arbitrary constant.*

Proof. See Appendix A.7. □

Please note that our analysis holds for any $\varepsilon\in(0,1)$. For the larger risk level $\varepsilon\in(0.5,1)$, the feasible region of CCP (9) can be non-convex (Henrion 2006) and intractable (see Proposition 12 in Appendix G). The two conditions in Theorem 4 may not be very strong and can be found in the CCP literature or relevant application problems. For example, proposition 5.1 of Van Ackooij and Malick 2019 studied the eventual convexity analysis of a CCP under an elliptical distribution and the condition that $\boldsymbol{u}=\boldsymbol{0}$ and $b_1(\boldsymbol{x})$ is a nonnegative constant, which is a special case of Condition (ii).

We also remark that either condition in Theorem 4 can be satisfied in practical application problems, for example, let us consider the following portfolio selection problem (Markowitz 1991, Pagnoncelli et al. 2009) as $v^*=\min_{\boldsymbol{x}\in\mathbb{R}^n_+}\{\boldsymbol{c}^\top\boldsymbol{x}\colon\mathbb{P}\{b_1\geq\tilde{\boldsymbol{\xi}}^\top\boldsymbol{x}\}\geq1-\varepsilon,\boldsymbol{e}^\top\boldsymbol{x}=1\}$, where decision vector $\boldsymbol{x}$ denotes the investment plan, scalar b_1 represents the portfolio return level, $\tilde{\boldsymbol{\xi}}$ is the stochastic return vector of n risky assets, and $\boldsymbol{c}$ is the cost vector. Suppose $\tilde{\boldsymbol{\xi}}$ follows a multi-variate elliptical distribution. Then using the notation in Section 2.4, the chance constrained portfolio selection problem is equivalent to

$$v^*=\min_{\boldsymbol{x}\in\mathbb{R}^n_+}\left\{\boldsymbol{c}^\top\boldsymbol{x}\colon b_1-\boldsymbol{\mu}^\top\boldsymbol{x}\geq\Phi^{-1}(1-\varepsilon)\sqrt{\boldsymbol{x}^\top\boldsymbol{\Sigma}\boldsymbol{x}},\boldsymbol{e}^\top\boldsymbol{x}=1\right\}. \tag{17}$$

Let us consider the following two cases of problem (17):

- To find an efficient portfolio, that is, to achieve the highest expected return of all the feasible portfolios with the same risk level (see the details in Fabozzi et al. 2012), the overall risk level $\sqrt{\boldsymbol{x}^\top \boldsymbol{\Sigma} \boldsymbol{x}}$ is a constant (i.e., Condition (i) of Theorem 4 is satisfied). In this case, according to Theorem 4, ALSO-X (10) returns an optimal solution; and
- When all the portfolios share the same expected return (see the discussions in Chow 1995), that is, when $\boldsymbol{u}^\top \boldsymbol{x}$ is a constant, Condition (ii) of Theorem 4 is satisfied. Similarly, according to Theorem 4, ALSO-X (10) returns an optimal solution.

Another byproduct of Theorem 4 is to demonstrate that ALSO-X can be strictly better than CVaR approximation. This is because CVaR approximation for the single linear CCP (9) under an elliptical distribution (see, e.g., theorem 9 in Chen and Xie 2019) can be written as

$$v^{\mathrm{CVaR}} = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon b_1(\boldsymbol{x}) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x}) \geq \left(\eta^{\mathrm{CVaR}} + \Phi^{-1}(1-\varepsilon)\right) \sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x})} \right\},$$

with constant

$$\eta^{\mathrm{CVaR}} = \frac{1}{\varepsilon} \int_{\frac{1}{2}(\Phi^{-1}(1-\varepsilon))^2}^{\infty} k\widehat{g}(z)dz - \Phi^{-1}(1-\varepsilon) = \overline{G}\left(\left(\Phi^{-1}(1-\varepsilon)\right)^2/2 \right)/\varepsilon - \Phi^{-1}(1-\varepsilon).$$

According to the proof of Theorem 4 in Appendix A.7, we observe that $\eta^{\mathrm{CVaR}} > 0$ for all $\varepsilon \in (0,1)$. Therefore, the feasible region of the CVaR approximation is strictly contained in that of CCP (9). As the ALSO-X is exact under the conditions of Theorem 4, ALSO-X can be strictly better than the CVaR approximation whenever $v^{\mathrm{CVaR}} > v^*$.

We also remark that albeit the first condition guarantees the optimality of ALSO-X (10), it might violate Assumption A2, namely, set $\mathcal{X}$ might not be convex. We show that these two conditions are the best that we can expect for the exactness of ALSO-X (10). In fact, Example 9 in Appendix B shows that ALSO-X (10) might not be able to find an optimal solution to CCP (9) if it violates both conditions of Theorem 4.

3.3. Approximation Ratio for the Covering CCPs under Discrete Support

In this subsection, we analyze the approximation ratio for a special case of finite-support CCP (8) with the covering structure (i.e., covering CCPs), where $\mathcal{X} = \mathbb{R}^n_+$, $\boldsymbol{c} \in \mathbb{R}^n_+$, the constraints $g(\boldsymbol{x}, \boldsymbol{\xi}^i) = \boldsymbol{b}^i - \boldsymbol{A}^i \boldsymbol{x}$ with $\boldsymbol{A}^i \in \mathbb{R}^{m\times n}_+$, $\boldsymbol{b}^i \in \mathbb{R}^m_{++}$, for all $i \in [N]$. Various applications have been studied in literature (see, e.g., Shiina 1999, Takyi and Lence 1999, Talluri et al. 2006, Deng and Shen 2016, Dentcheva et al. 2000, Xie and Ahmed 2020, Qiu et al. 2014) that can be formulated as covering CCPs. Without loss of generality, $\boldsymbol{b}^i$ can always be normalized to $\boldsymbol{e}$, and thus covering CCP (1) admits the following form

$$v^* = \min_{\boldsymbol{x}\in\mathbb{R}^n_+, \boldsymbol{z}\in\{0,1\}^N} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \sum_{i\in[N]} z_i \geq N - \lfloor N\varepsilon \rfloor, \boldsymbol{A}^i \boldsymbol{x} \geq z_i \mathbf{e}, \forall i \in [N] \right\}, \tag{18}$$

and the corresponding ALSO-X (7) is

$$v^A = \min_{t} \quad t, \tag{19a}$$

$$\text{s.t.} \quad (\boldsymbol{x}^*, \boldsymbol{s}^*) \in \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^n_+, \boldsymbol{s}\in\mathbb{R}^N_+} \left\{ \frac{1}{N} \sum_{i\in[N]} s_i \colon \boldsymbol{c}^\top \boldsymbol{x} \le t, \boldsymbol{A}^i \boldsymbol{x} \ge \boldsymbol{e} - s_i \boldsymbol{e}, \forall i \in [N] \right\}, \tag{19b}$$

$$\sum_{i\in[N]} \mathbb{I}(s_i^* = 0) \ge N - \lfloor N\varepsilon \rfloor. \tag{19c}$$

Consider the following continuous relaxation of the covering CCP (18) as

$$v^{rel} = \min_{\boldsymbol{x}\in\mathbb{R}^n_+, \boldsymbol{s}\in\mathbb{R}^N_+} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \sum_{i\in[N]} s_i \le \lfloor N\varepsilon \rfloor, \boldsymbol{A}^i \boldsymbol{x} \ge \boldsymbol{e} - s_i \boldsymbol{e}, \forall i \in [N] \right\}. \tag{20}$$

Particularly, we observe that (i) in the continuous relaxation (18), at optimality, we must have $s_i \in [0,1]$ for all $i \in [N]$; and (ii) the continuous relaxation value v^{rel} is a lower bound for the covering CCP (18). Next, we show that for any $t \ge (\lfloor N\varepsilon \rfloor + 1)v^{rel}$, any optimal solution to the hinge-loss approximation (19b) is feasible to the covering CCP (18). This implies that $v^A/v^* \le v^A/v^{rel} \le \lfloor N\varepsilon \rfloor + 1$, i.e., ALSO-X (19) yields a $(\lfloor N\varepsilon \rfloor + 1)-$approximation guarantee, achieving the same best known approximation ratio for the finite-support covering CCPs (Ahmed and Xie 2018). When employing ALSO-X Algorithm 1, if finding a feasible solution of a covering CCP is difficult, one can use $(\lfloor N\varepsilon \rfloor + 1)v^{rel}$ as a valid upper bound. In the numerical study, we apply this strategy when running ALSO-X Algorithm 1 to solve a covering CCP instance.

Theorem 5 *For the covering CCP* (18)*, the* ALSO-X (19) *yields a* $(\lfloor N\varepsilon \rfloor + 1)-$*approximation guarantee, that is,* $v^A/v^* \le \lfloor N\varepsilon \rfloor + 1$.

Proof. See Appendix A.8. □

Finally, we conclude this section by showing that the approximation ratio of ALSO-X (19) for the covering CCP (18) is tight.

Proposition 5 *For the covering CCP* (18)*, the* $(\lfloor N\varepsilon \rfloor + 1)-$*approximation ratio of* ALSO-X (19) *is tight, i.e., it is possible that* $v^A/v^* = \lfloor N\varepsilon \rfloor + 1$.

Proof. See Appendix A.9. □

4. ALSO-X+ Algorithm: Breaking the Symmetry and Improving ALSO-X Algorithm 1 using Alternating Minimization Method

In Section 2.2, recall that we derive ALSO-X (7) and its related Algorithm 1 by letting the functional variable $z(\boldsymbol{\xi}) = 1$ in the CCP (5). As shown in the previous sections, ALSO-X (7) successfully

provides exact optimal and approximate solutions for many special families of CCP (5). On the other hand, simply forcing the functional variable $z(\boldsymbol{\xi}) = 1$ in CCP (5) might not be ideal, for instance, Example 3 demonstrates that ALSO-X (7) might be fooled if the random parameters $\tilde{\boldsymbol{\xi}}$ obey a joint symmetric distribution. This motivates us to improve the ALSO-X Algorithm 1 by optimizing the functional variable $z(\cdot)$ as well. In general, optimizing over both $s(\cdot)$ and $z(\cdot)$ can be difficult. Thus, we propose to enhance ALSO-X Algorithm 1, termed "ALSO-X+ algorithm," by using a better Alternating Minimization (AM) method, which is to optimize over $(\boldsymbol{x}, s(\cdot))$ and $z(\cdot)$ of the lower-level problem (5b) in an alternating fashion. The key idea of ALSO-X+ algorithm is that in Step 4 of ALSO-X Algorithm 1, one should run the AM method if the current solution is infeasible (i.e., the optimal solution $(\boldsymbol{x}^*, s^*(\cdot))$ of the hinge-loss approximation (7b) violates the chance constraint, $\mathbb{P}\{\tilde{\boldsymbol{\xi}}\colon s^*(\tilde{\boldsymbol{\xi}}) > 0\} > \varepsilon$).

4.1. The Proposed Alternating Minimization (AM) Method

To begin with, we first introduce the AM method. As mentioned earlier, for a given t, when the hinge-loss approximation (7b) is unable to provide a feasible solution to the CCP (1), i.e., its optimal solution $(\boldsymbol{x}^*, s^*(\cdot))$ is subject to $\mathbb{P}\{\tilde{\boldsymbol{\xi}}\colon s^*(\tilde{\boldsymbol{\xi}}) > 0\} > \varepsilon$. Under this circumstance, we run the AM method to optimize the lower-level problem (5b) in hope of overcoming the infeasibility, which proceeds as follows. First, we observe that the constraint system in the lower-level problem (5b) can be separated into two parts, with respect to the functional variable $z(\cdot)$ and with respect to variables $\boldsymbol{x}$ and $s(\cdot)$. This allows us to optimize over $z(\cdot)$ and $(\boldsymbol{x}, s(\cdot))$ in an iterative way. Specifically, at iteration $k+1$, when fixing $(\boldsymbol{x}, s(\cdot)) = (\boldsymbol{x}^k, s^k(\cdot))$ with $(\boldsymbol{x}^k, s^k(\cdot))$ from the previous iteration in the lower-level problem (5b), we solve the following optimization problem:

$$z^{k+1}(\cdot) \in \arg\min_{z(\cdot)} \left\{ \mathbb{E}\left[z(\tilde{\boldsymbol{\xi}}) s^k(\tilde{\boldsymbol{\xi}}) \right] : z(\tilde{\boldsymbol{\xi}}) \in [0,1], \mathbb{E}[z(\tilde{\boldsymbol{\xi}})] \geq 1 - \varepsilon \right\}, \tag{21a}$$

which can be done via sorting the values of $\{s^k(\boldsymbol{\xi})\}_{\boldsymbol{\xi}\in\Xi}$. Next fixing the value of the functional variable $z(\cdot) = z^{k+1}(\cdot)$, we solve the following convex optimization problem:

$$\left(\boldsymbol{x}^{k+1}, s^{k+1}(\cdot)\right) \in \arg\min_{\boldsymbol{x}\in\mathcal{X}, s(\cdot)} \left\{ \mathbb{E}\left[z^{k+1}(\tilde{\boldsymbol{\xi}}) s(\tilde{\boldsymbol{\xi}}) \right] : \boldsymbol{c}^\top \boldsymbol{x} \leq t, g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq s(\tilde{\boldsymbol{\xi}}), s(\tilde{\boldsymbol{\xi}}) \geq 0 \right\}. \tag{21b}$$

Note that the problem (21b) can be further simplified as

$$\boldsymbol{x}^{k+1} \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \mathbb{E}\left[z^{k+1}(\tilde{\boldsymbol{\xi}}) [g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]_+ \right] : \boldsymbol{c}^\top \boldsymbol{x} \leq t \right\},$$

which can be solved using the subgradient descent method proposed in Section 2.3 for instance. We continue this procedure until reaching the stopping criterion. The detailed implementation can be found in Algorithm 3. Note that according to bilinear programming Formulation (3), the output solution $\boldsymbol{x}^{k+1}$ is feasible to CCP (1) if and only if $\mathbb{E}[z^{k+1}(\tilde{\boldsymbol{\xi}}) s^{k+1}(\tilde{\boldsymbol{\xi}})] = 0$.

Algorithm 3 Alternating Minimization (AM) Method to Solve the Lower-level Problem (5b)

1: Let $k=0$. Let δ_2 denote the stopping tolerance parameter, t be the current given value of the upper-level problem, and $z^0(\cdot)$ be the given initial solution of $z(\cdot)$, respectively
2: **do**
3: Solve (21a) and (21b) with optimal solutions $z^{k+1}(\cdot)$ and $(\boldsymbol{x}^{k+1}, s^{k+1}(\cdot))$, respectively
4: Let $\Delta = \left|\mathbb{E}[z^{k+1}(\tilde{\boldsymbol{\xi}})s^{k+1}(\tilde{\boldsymbol{\xi}})] - \mathbb{E}[z^{k}(\tilde{\boldsymbol{\xi}})s^{k}(\tilde{\boldsymbol{\xi}})]\right|$ and $k=k+1$
5: **while** $\Delta \geq \delta_2$
6: The output solution $\boldsymbol{x}^{k+1}$ is feasible to CCP (1) if $\mathbb{E}[z^{k+1}(\tilde{\boldsymbol{\xi}})s^{k+1}(\tilde{\boldsymbol{\xi}})]=0$; otherwise, it is infeasible

In AM Algorithm 3, we observe that the sequence of objective values $\{\mathbb{E}[z^k(\tilde{\boldsymbol{\xi}})s^k(\tilde{\boldsymbol{\xi}})]\}_{k\in\mathbb{Z}_+}$ converges. This demonstrates that the stopping criterion using the objective values is indeed valid.

Proposition 6 *The sequence of objective values* $\{\mathbb{E}[z^k(\tilde{\boldsymbol{\xi}})s^k(\tilde{\boldsymbol{\xi}})]\}_{k\in\mathbb{Z}_+}$ *generated by the* AM *Algorithm 3 is monotonically nonincreasing, bounded from below, and hence converges.*

Proof. See Appendix A.10. □

It is worthy of mentioning that the sequence of solutions $\{(\boldsymbol{x}^k, s^k(\cdot), z^k(\cdot))\}_{k\in\mathbb{Z}_+}$ generated by the AM Algorithm 3 might not converge. If this case occurs, we choose a convergent subsequence of $\{(\boldsymbol{x}^k, s^k(\cdot), z^k(\cdot))\}_{k\in\mathbb{Z}_+}$ and its accumulative point as the output.

4.2. AM Method is Better Than Difference-of-Convex (DC) Approach

This subsection compares the AM method with the well-known difference-of-convex (DC) approach to solve the lower-level problem (5b), and shows that the AM method provides a better-quality solution than that of the DC approach.

We first apply the well-known difference-of-convex (DC) approach (see, e.g., section 2 of Tao and An 1997) to solve the lower-level problem (5b). Note that this DC approach is different compared to the one studied in Hong et al. (2011), where the latter directly approximated the chance constraint using difference-of-convex functions. Observe that the objective function in (5b) can be rewritten as the difference of two convex quadratic functions:

$$\mathbb{E}\left[z(\tilde{\boldsymbol{\xi}})s(\tilde{\boldsymbol{\xi}})\right] = \frac{1}{4}\mathbb{E}\left[\left(z(\tilde{\boldsymbol{\xi}})+s(\tilde{\boldsymbol{\xi}})\right)^2 - \left(z(\tilde{\boldsymbol{\xi}})-s(\tilde{\boldsymbol{\xi}})\right)^2\right].$$

Next, the DC approach proceeds as follows. At iteration $k+1$, we replace $(z(\boldsymbol{\xi})-s(\boldsymbol{\xi}))^2$ by its first order Taylor approximation using the solution from the previous iteration, i.e.,

$$(z(\boldsymbol{\xi})-s(\boldsymbol{\xi}))^2 \approx \left(z^k(\boldsymbol{\xi})-s^k(\boldsymbol{\xi})\right)^2 + 2z^k(\boldsymbol{\xi})\left(z(\boldsymbol{\xi})-z^k(\boldsymbol{\xi})\right) - 2s^k(\boldsymbol{\xi})\left(s(\boldsymbol{\xi})-s^k(\boldsymbol{\xi})\right),$$

and solve the following convex program:

$$\begin{aligned}\left(\boldsymbol{x}^{k+1}, s^{k+1}(\cdot), z^{k+1}(\cdot)\right) \in \underset{\boldsymbol{x}\in\mathcal{X}, s(\cdot), z(\cdot)}{\arg\min} \Bigg\{ & \frac{1}{4}\mathbb{E}\left[\left(z(\tilde{\boldsymbol{\xi}}) + s(\tilde{\boldsymbol{\xi}})\right)^2\right] - \frac{1}{4}\mathbb{E}\left[\left(z^k(\tilde{\boldsymbol{\xi}}) - s^k(\tilde{\boldsymbol{\xi}})\right)^2\right] \\ & - \frac{1}{4}\mathbb{E}\left[2z^k(\tilde{\boldsymbol{\xi}})\left(z(\tilde{\boldsymbol{\xi}}) - z^k(\tilde{\boldsymbol{\xi}})\right) - 2s^k(\tilde{\boldsymbol{\xi}})\left(s(\tilde{\boldsymbol{\xi}}) - s^k(\tilde{\boldsymbol{\xi}})\right)\right] : \\ & \boldsymbol{c}^\top \boldsymbol{x} \le t, g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \le s(\tilde{\boldsymbol{\xi}}), s(\tilde{\boldsymbol{\xi}}) \ge 0, z(\tilde{\boldsymbol{\xi}}) \in [0,1] \Bigg\}. \end{aligned} \tag{22}$$

And repeat this process until the objective functions $\{\mathbb{E}[z^k(\tilde{\boldsymbol{\xi}})s^k(\tilde{\boldsymbol{\xi}})]\}_{k\in\mathbb{Z}_+}$ converge.

Since both AM method and DC approach find a stationary point, the formal comparison between the AM method and the DC approach relies on their stationary conditions. Particularly, the AM method generates a stationary point $(\boldsymbol{x}^{\mathrm{AM}}, s^{\mathrm{AM}}(\cdot), z^{\mathrm{AM}}(\cdot)))$ that solves problem (21a) and (21b) when $(\boldsymbol{x}^{k+1}, s^{k+1}(\cdot)) = (\boldsymbol{x}^{\mathrm{AM}}, s^{\mathrm{AM}}(\cdot))$ and $z^k(\cdot) = z^{\mathrm{AM}}(\cdot)$ if and only if its satisfies the following stationary condition:

$$\begin{aligned} & \mathbb{E}\left[s^{\mathrm{AM}}(\cdot)\left[z(\cdot) - z^{\mathrm{AM}}(\cdot)\right]\right] \ge 0, \mathbb{E}\left[z^{\mathrm{AM}}(\cdot)\left[s(\cdot) - s^{\mathrm{AM}}(\cdot)\right]\right] \ge 0, \\ & \forall\, (\boldsymbol{x}, s(\cdot), z(\cdot)) \text{ satisfies the constraints in the lower-level problem (5b).} \end{aligned} \tag{23}$$

On the other hand, the DC method generates a stationary point $(\boldsymbol{x}^{\mathrm{DC}}, s^{\mathrm{DC}}(\cdot), z^{\mathrm{DC}}(\cdot))$ that solves (22) when $(\boldsymbol{x}^k, s^k(\cdot), z^k(\cdot)) = (\boldsymbol{x}^{\mathrm{DC}}, s^{\mathrm{DC}}(\cdot), z^{\mathrm{DC}}(\cdot))$ if and only if its satisfies the following stationary condition:

$$\begin{aligned} & \mathbb{E}\left[s^{\mathrm{DC}}(\cdot)\left[z(\cdot) - z^{\mathrm{DC}}(\cdot)\right]\right] + \mathbb{E}\left[z^{\mathrm{DC}}(\cdot)\left[s(\cdot) - s^{\mathrm{DC}}(\cdot)\right]\right] \ge 0, \\ & \forall\, (\boldsymbol{x}, s(\cdot), z(\cdot)) \text{ satisfies the constraints in the lower-level problem (5b).} \end{aligned} \tag{24}$$

Note that the set of the stationary points satisfying the condition (23) of the AM method is contained in that satisfying the condition (24) of the DC approach. This concludes that the AM method is better than the DC approach.

Proposition 7 *Given t, when solving the lower-level problem* (5b)*, the* AM *method can find a better solution than that of* DC *approach.*

Then following example demonstrates that the solution from AM method can be indeed strictly better than that of the DC approach.

Example 5 Let us revisit Example 3. Given $t = 0.5$, then the lower-level problem (5b) admits the following form

$$\min_{\boldsymbol{x}\in\mathbb{R}^2_+, \boldsymbol{s}\in\mathbb{R}^3_+, \boldsymbol{z}} \left\{ \frac{1}{3}\sum_{i\in[3]} z_i s_i : \begin{array}{l} 2x_1 + 3x_2 \ge 1 - s_1, 2x_1 + x_2 \ge 1 - s_2 x_1 + 2x_2 \ge 1 - s_3, \\ \sum_{i\in[3]} z_i \ge 2, \boldsymbol{z} \in [0,1]^3, x_1 + x_2 \le 0.5 \end{array} \right\}. \tag{25a}$$

When running AM Algorithm 3 with an initial solution $\boldsymbol{z}^0 = [1,1,1]$, the stationary solution is $\boldsymbol{z}^{\mathrm{AM}} = [1,0,1]$, $\boldsymbol{s}^{\mathrm{AM}} = [0, 0.5, 0]$, $\boldsymbol{x}^{\mathrm{AM}} = [0, 0.5]$. When using DC approach with an initial solution

$\boldsymbol{z}^0 = [1,1,1]$, $\boldsymbol{s}^0 = [1,1,1]$, the stationary solution is $\boldsymbol{z}^{\rm DC} = [0,1,1]$, $\boldsymbol{s}^{\rm DC} = [0,0.25,0.25]$, $\boldsymbol{x}^{\rm DC} = [0.25,0.25]$. More importantly, since $\sum_{i\in[3]} \mathbb{I}(s_i^{\rm AM} = 0) = 2 \geq 2$ and $\sum_{i\in[3]} \mathbb{I}(s_i^{\rm DC} = 0) = 1 < 2$, the AM method is able to find a feasible solution to the CCP for this example, while the DC approach is unable to. ◇

4.3. ALSO-X+ Algorithm

This subsection integrates ALSO-X Algorithm 1 with the AM Algorithm 3 as ALSO-X+ algorithm, to improve the performance of ALSO-X Algorithm 1. In ALSO-X+ algorithm, we first execute ALSO-X Algorithm 1, and when Step 4 of ALSO-X Algorithm 1 encounters an infeasible solution (i.e., the optimal solution $(\boldsymbol{x}^*, s^*(\cdot))$ of the hinge-loss approximation (7b) violates the chance constraint $\mathbb{P}\{\tilde{\boldsymbol{\xi}}\colon s^*(\tilde{\boldsymbol{\xi}}) > 0\} > \varepsilon$), then we run the AM Algorithm 3 with the same t and see if we are able to find a feasible solution. If YES, we further decrease $t_U = t$; otherwise, we increase $t_L = t$. The detailed procedure for the ALSO-X+ algorithm is shown in Algorithm 4.

Algorithm 4 The Proposed ALSO-X+ Algorithm

1: **Input:** Let δ_1 denote the stopping tolerance parameter, t_L and t_U be the known lower and upper bounds of the optimal value of CCP (1), respectively
2: **while** $t_U - t_L > \delta_1$ **do**
3: Let $t = (t_L + t_U)/2$ and $(\boldsymbol{x}^*, s(\cdot))$ be an optimal solution of the hinge-loss approximation (7b)
4: Let $t_U = t$ if $\mathbb{P}\{\tilde{\boldsymbol{\xi}}\colon s^*(\tilde{\boldsymbol{\xi}}) > 0\} \leq 1 - \varepsilon$; otherwise, run the AM Algorithm 3. If the solution output from the AM Algorithm 3 is feasible to the CCP, let $t_U = t$; otherwise, $t_L = t$
5: **end while**
6: **Output:** A feasible solution $\boldsymbol{x}^*$ and its objective value to CCP (1)

We make the following remarks about ALSO-X+ Algorithm 4.

(i) We can use the solutions of the hinge-loss approximation (7b) as warm-starts for the AM Algorithm 3;

(ii) For the linear CCP, i.e., $g_i(\boldsymbol{x}, \boldsymbol{\xi}) = \boldsymbol{\xi}^\top \boldsymbol{a}_i(\boldsymbol{x}) - b_i(\boldsymbol{x})$ with affine functions $\boldsymbol{a}_i(\boldsymbol{x}), b_i(\boldsymbol{x})$ and set $\mathcal{X}$ is a polyhedron. Similar to ALSO-X Algorithm 1, we can use parametric linear programming (Adler and Monteiro 1992) to decrease the number of bisection needed. That is, after Step 3, we can record the lower and upper bounds of the allowable range of the value t. Then at Step 4, if the current solution is not feasible to CCP (1), we can update t to be the upper bound of its allowable range; otherwise, we can let t be equal to the lower bound of the allowable range. We then continue the binary search procedure;

(iii) Inherited from ALSO-X Algorithm 1, ALSO-X+ Algorithm 4 always provides a better solution than that of the CVaR approximation given that Assumptions A1-A2 hold and the tolerance $\delta_1 = 0$; and

(iv) For the nonconvex set $\mathcal{X}$, Example 2 can be further used to demonstrate that the CVaR approximation can also outperform ALSO-X+ Algorithm 4. That is, in Example 2, the solution output from ALSO-X+ Algorithm 4 is the same as that from ALSO-X, which is not optimal, while the CVaR approximation provides the optimal solution.

Besides, incorporating AM Algorithm 3 in ALSO-X+ Algorithm 4 helps break the symmetry in the hinge-loss approximation (7b) by assigning different weights to the violations of uncertain constraints. Specifically, in AM Algorithm 3, when fixing $z(\cdot) = z^k(\cdot)$ with $\mathbb{E}[z^k(\tilde{\boldsymbol{\xi}})] = 1 - \varepsilon$, the problem (21b) tends to focus on the $1 - \varepsilon$ portion of uncertain constraints rather than using all of them in the hinge-loss approximation (7b). In the following example, we show that due to the symmetry of the random parameters, ALSO-X (7) is unable to provide an optimal solution, while ALSO-X+ Algorithm 4 with the tolerance $\delta_1 = 0$ can.

Example 6 Let us revisit Example 3. Suppose that $t = 0.6$, then the optimal solution provided by ALSO-X (7) is $x_1^* = x_2^* = 0.3$, $s_1^* = 0$, $s_2^* = s_3^* = 0.1$, which violates the chance constraint. Invoking the AM Algorithm 3 with initial $s_i^0 = s_i^*$ for each $i \in [3]$, at the second iteration of the AM Algorithm 3, we find an optimal solution $x_1^2 = 0.4, x_2^2 = 0.2$, $s_1^2 = s_2^2 = 0$, $s_3^2 = 0.2$, and $z_1^2 = z_2^2 = 1$, $z_3^2 = 0$ to the lower-level problem (5b) with $t = 0.6$. Thus, if $t = 0.6$, ALSO-X+ Algorithm 4 further reduce the $t_U = t = 0.6$. In fact, in this example, ALSO-X+ Algorithm 4 with the tolerance $\delta_1 = 0$ finds the optimal solution of the CCP. ◇

Although improving ALSO-X (7), ALSO-X+ Algorithm 4 might not always be able to find an optimal solution of CCP (1), as illustrated in Example 10 of Appendix B. Interested readers are referred to Appendix C for an illustration of comparisons among ALSO-X, CVaR approximation, and ALSO-X+ Algorithm.

5. Extension to Distributionally Robust Chance Constrained Programs (DRCCP) with Wasserstein Distance

In practice, the distributional information of random parameters $\tilde{\boldsymbol{\xi}}$ might not be fully known, making it difficult to commit to a single known probability distribution $\mathbb{P}$. Under this circumstance, to hedge against distributional ambiguity, we consider the distributionally robust chance constrained programs (DRCCPs), which require the chance constraint to be satisfied for all the probability distributions from a family of distributions, termed "ambiguity set." That is, following many recent

works in DRCCP (Xie and Ahmed 2020, Xie 2019, Chen et al. 2018, Ji and Lejeune 2020, Chen and Xie 2019), we consider the DRCCP of the form

$$\min_{\boldsymbol{x}\in\mathcal{X}}\left\{\boldsymbol{c}^\top\boldsymbol{x}\colon \inf_{\mathbb{P}\in\mathcal{P}}\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon g(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\le 0\right\}\ge 1-\varepsilon\right\}, \tag{26}$$

where ambiguity set $\mathcal{P}$ denotes a subset of probability distributions $\mathbb{P}$ defined on $(\Omega,\mathcal{F})$ and induced by the random parameters $\tilde{\boldsymbol{\xi}}$, and risk level $\varepsilon\in(0,1)$. Specifically, we study the DRCCP under $\infty-$Wasserstein ambiguity set (see, e.g., Bertsimas et al. 2018, Xie 2020). The $q-$Wasserstein ambiguity set is defined as

$$\mathcal{P}_q^W=\left\{\mathbb{P}\colon \mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\in\Xi\right\}=1, W_q(\mathbb{P},\mathbb{P}_{\tilde{\zeta}})\le\theta\right\},$$

where for any $q\in[1,\infty]$, the $q-$Wasserstein distance is defined as

$$W_q(\mathbb{P}_1,\mathbb{P}_2)=\inf\left\{\left[\int_{\Xi\times\Xi}\|\boldsymbol{\xi}_1-\boldsymbol{\xi}_2\|^q\,\mathbb{Q}(d\boldsymbol{\xi}_1,d\boldsymbol{\xi}_2)\right]^{\frac{1}{q}}\colon \begin{array}{l}\mathbb{Q}\text{ is a joint distribution of }\tilde{\boldsymbol{\xi}}_1\text{ and }\tilde{\boldsymbol{\xi}}_2\\ \text{with marginals }\mathbb{P}_1\text{ and }\mathbb{P}_2,\text{ respectively}\end{array}\right\},$$

where $\theta\ge 0$ is the Wasserstein radius, and $\mathbb{P}_{\tilde{\zeta}}$ denotes the reference distribution induced by random parameters $\tilde{\boldsymbol{\zeta}}$. For example, $\mathbb{P}_{\tilde{\zeta}}$ can be an empirical distribution with $\tilde{\boldsymbol{\zeta}}$ being a uniform discrete random vector. Note that if $q=\infty$, the $\infty-$Wasserstein distance is reduced to

$$W_\infty(\mathbb{P}_1,\mathbb{P}_2)=\inf\left\{\operatorname{ess.sup}\|\boldsymbol{\xi}_1-\boldsymbol{\xi}_2\|\,\mathbb{Q}(\mathrm{d}\boldsymbol{\xi}_1,\mathrm{d}\boldsymbol{\xi}_2)\colon \begin{array}{l}\mathbb{Q}\text{ is a joint distribution of }\tilde{\boldsymbol{\xi}}_1\text{ and }\tilde{\boldsymbol{\xi}}_2\\ \text{with marginals }\mathbb{P}_1\text{ and }\mathbb{P}_2,\text{ respectively}\end{array}\right\}.$$

Throughout this section, we assume that

A3 The reference distribution $\mathbb{P}_{\tilde{\zeta}}$ is sub-Gaussian, that is, $\mathbb{P}_{\tilde{\zeta}}\{\tilde{\boldsymbol{\zeta}}:\|\tilde{\boldsymbol{\zeta}}\|\ge t\}\le C_1\exp(-C_2t^2)$ for some positive constants C_1,C_2.

It is worthy of noting that the sub-Gaussian assumption ensures the weak compactness of $\infty-$Wasserstein ambiguity set and thus ensures the strong duality of reformulating the worst-case expectation under $\infty-$Wasserstein ambiguity set. Particularly, this paper mainly focuses on empirical or elliptical reference distributions, which clearly satisfy Assumption A3.

Under this setting, DRCCP (26) can be written as

$$v_\infty^*:=\min_{\boldsymbol{x}\in\mathcal{X}}\left\{\boldsymbol{c}^\top\boldsymbol{x}\colon \inf_{\mathbb{P}\in\mathcal{P}_\infty^W}\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon g_i(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\le 0,\forall i\in[I]\right\}\ge 1-\varepsilon\right\}. \tag{27}$$

It turns out that DRCCP (27) admits a neat equivalent representation.

Proposition 8 *Under $\infty-$Wasserstein ambiguity set, DRCCP* (27) *is equivalent to*

$$v_\infty^*=\min_{\boldsymbol{x}\in\mathcal{X}}\left\{\boldsymbol{c}^\top\boldsymbol{x}\colon \mathbb{P}_{\tilde{\zeta}}\left\{\tilde{\boldsymbol{\zeta}}\colon \bar{g}_i(\boldsymbol{x},\tilde{\boldsymbol{\zeta}})\le 0,\forall i\in[I]\right\}\ge 1-\varepsilon\right\}, \tag{28}$$

where the convex and lower semi-continuous function $\bar{g}_i:\mathbb{R}^n\times\Xi\to\mathbb{R}$ is defined as $\bar{g}_i(\boldsymbol{x},\boldsymbol{\zeta}):=\max_{\boldsymbol{\xi}}\{g_i(\boldsymbol{x},\boldsymbol{\xi})\colon\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|\le\theta\}$ for each $i\in[I]$.

Proof. See Appendix A.11. □

The reformulation in Proposition 8 implies that DRCCP (27) under $\infty-$Wasserstein ambiguity set is equivalent to a regular CCP (28). In fact, we anticipate that the worst-case ALSO-X and the worst-case CVaR approximation under $\infty-$Wasserstein ambiguity set are equivalent to directly applying ALSO-X and CVaR approximation to solve CCP (28). This observation motivates us to show that under $\infty-$Wasserstein ambiguity set, the worst-case ALSO-X is better than the worst-case CVaR approximation, which is detailed in the next subsection.

5.1. The Worst-case ALSO-X Outperforms the Worst-case CVaR Approximation

In this subsection, we introduce the notions of the worst-case ALSO-X and the worst-case CVaR approximation, and then demonstrate that the worst-case ALSO-X always outperforms the worst-case CVaR approximation under $\infty-$Wasserstein ambiguity set.

Similar to ALSO-X (7), we derive the worst-case ALSO-X counterpart under $\infty-$Wasserstein ambiguity set. That is, in the worst-case ALSO-X, we first solve the worst-case hinge-loss approximation, which is to minimize the least-favorable expectation of function $\max_{i\in[I]} g_i(\boldsymbol{x},\tilde{\boldsymbol{\xi}})_+$, i.e., the objective function of (29b) is to minimize the worst-case objective function of the hinge-loss approximation (7b), and check if its optimal solution $\boldsymbol{x}^*$ satisfies the distributionally robust chance constraint in (27) or not. If the answer is YES, we continue reducing the upper bound of the objective value t, and otherwise, we increase t. This procedure can be formally formulated as

$$v_\infty^A = \min_t \quad t, \tag{29a}$$

$$\text{s.t.} \quad \boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}\in\mathcal{P}_\infty^W} \left\{ \mathbb{E}_{\mathbb{P}}\left[\max_{i\in[I]} g_i(\boldsymbol{x},\tilde{\boldsymbol{\xi}})_+\right] : \boldsymbol{c}^\top\boldsymbol{x}\le t \right\}, \tag{29b}$$

$$\inf_{\mathbb{P}\in\mathcal{P}_\infty^W} \mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon g_i(\boldsymbol{x}^*,\tilde{\boldsymbol{\xi}})\le 0, \forall i\in[I]\right\} \ge 1-\varepsilon. \tag{29c}$$

For DRCCP (27), the worst-case CVaR approximation is defined as

$$v_\infty^{\mathrm{CVaR}} = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top\boldsymbol{x}\colon \sup_{\mathbb{P}\in\mathcal{P}_\infty^W} \inf_\beta \left[\beta + \frac{1}{\varepsilon}\mathbb{E}_{\mathbb{P}}\left[\max_{i\in[I]}\left(g_i(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-\beta\right)_+\right]\right]\le 0 \right\}. \tag{30}$$

The next proposition formally derives the equivalent reformations of the worst-case ALSO-X (29) and the worst-case CVaR approximation (30), respectively.

Proposition 9 *Under $\infty-$Wasserstein ambiguity set, we have*

(i) the worst-case ALSO-X (29) *is equivalent to*

$$\begin{aligned} v_\infty^A = \min_t \quad & t, \\ \text{s.t.} \quad & \boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \mathbb{E}_{\mathbb{P}}\left[\max_{i\in[I]} \bar{g}_i(\boldsymbol{x},\tilde{\boldsymbol{\xi}})_+\right] : \boldsymbol{c}^\top\boldsymbol{x}\le t \right\}, \\ & \mathbb{P}_{\tilde{\zeta}}\left\{\tilde{\boldsymbol{\zeta}}\colon \bar{g}_i(\boldsymbol{x}^*,\tilde{\boldsymbol{\zeta}})\le 0, \forall i\in[I]\right\} \ge 1-\varepsilon; \end{aligned} \tag{31}$$

(ii) the worst-case CVaR *approximation* (30) *is equivalent to*

$$v_\infty^{\mathrm{CVaR}} = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \min_{\beta} \left[\beta + \frac{1}{\varepsilon} \mathbb{E}_{\mathbb{P}_{\tilde{\boldsymbol{\zeta}}}} \left\{ \max_{i\in[I]} \left\{ \bar{g}_i(\boldsymbol{x}, \tilde{\boldsymbol{\zeta}}) \right\} - \beta \right\}_+ \right] \le 0 \right\}. \tag{32}$$

Proof. See Appendix A.12. □

We remark that both the worst-case ALSO-X (31) and the worst-case CVaR approximation (30) under $\infty-$Wasserstein ambiguity set can be interpreted as applying ALSO-X and CVaR approximation of the regular CCP (28), respectively. Therefore, the results in previous sections hold for DRCCP (28) including that we can simply apply ALSO-X+ to CCP (28). More importantly, following the spirit of Section 3.1, we can conclude that the worst-case ALSO-X is better than the worst-case CVaR approximation under $\infty-$Wasserstein ambiguity set.

Theorem 6 *For DRCCP with $\infty-$Wasserstein ambiguity set, the worst-case* ALSO-X *outperforms the worst-case* CVaR *approximation.*

Proof. See Appendix A.13. □

We also remark that under some additional assumptions of the functions $\{g_i(\cdot,\cdot)\}$, their robust counterparts $\{\bar{g}_i(\cdot,\cdot)\}$ admit simple representations. Interested readers are referred to the work (Ben-Tal et al. 2009) for many different choices of functions $\{g_i(\cdot,\cdot)\}$. Below, we list two classes of functions:

(i) When the functions are bi-affine, i.e., $g_i(\boldsymbol{x},\boldsymbol{\xi}) = \boldsymbol{\xi}^\top \boldsymbol{a}_i(\boldsymbol{x}) - b_i(\boldsymbol{x})$ with affine functions $\boldsymbol{a}_i(\boldsymbol{x}), b_i(\boldsymbol{x})$ for each $i \in [I]$, we have

$$\bar{g}_i(\boldsymbol{x},\boldsymbol{\zeta}) = \theta \left\| \boldsymbol{a}_i^\top(\boldsymbol{x}) \right\|_* + \tilde{\boldsymbol{\zeta}}^\top \boldsymbol{a}_i(\boldsymbol{x}) - b_i(\boldsymbol{x}), \forall i \in [I].$$

Note that the bi-affinity assumption of $\{g_i(\cdot,\cdot)\}_{i\in[I]}$ has been commonly used in many DRCCP literature (see, e.g., Hanasusanto et al. 2015, 2017, Xie and Ahmed 2018a, Xie 2019).

(ii) When the norm is L_∞ (i.e., $\|\cdot\| = \|\cdot\|_\infty$) and the function $g_i(\boldsymbol{x},\boldsymbol{\xi})$ is monotone non-decreasing in $\boldsymbol{\xi}$ for any $\boldsymbol{x} \in \mathcal{X}$ and $i \in [I]$, we have

$$\bar{g}_i(\boldsymbol{x},\boldsymbol{\zeta}) = g_i(\boldsymbol{x}, \boldsymbol{\zeta} + \theta \boldsymbol{e}), \forall i \in [I].$$

This monotonicity structure has been studied in the recent works (Zhang et al. 2021, Xie 2020, Chen and Xie 2020).

5.2. Exactness of the Worst-case ALSO-X

Similar to Theorem 4, we are able to identify two sufficient conditions under which the worst-case ALSO-X (29) can provide an optimal solution to DRCCP (27). Particularly, we consider the single DRCCP and elliptical reference distribution with the following condition.

Proposition 10 *Suppose that the reference distribution $\mathbb{P}_{\tilde{\zeta}}$ is elliptical $\mathbb{P}_{\mathrm{E}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{g})$, and the norm defining the Wasserstein distance is the Mahalanobis norm associated with the positive definite matrix $\boldsymbol{\Sigma}$, i.e., $\|\boldsymbol{y}\| = \sqrt{\boldsymbol{y}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}}$, for some $\boldsymbol{y} \in \mathbb{R}^n$. If $I = 1$ and the random function $g_1(\boldsymbol{x}, \boldsymbol{\xi}) = \boldsymbol{\xi}^\top \boldsymbol{a}_1(\boldsymbol{x}) - b_1(\boldsymbol{x})$, then the worst-case* ALSO-X (29) *provides an optimal solution to DRCCP* (27) *under $\infty-$Wasserstein ambiguity set if (i) $\mathcal{X} \subseteq \{\boldsymbol{x} : \sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x})} = C\}$, where C is a positive constant; or (ii) $\mathcal{X} \subseteq \{\boldsymbol{x} : b_1(\boldsymbol{x}) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x}) = C\}$, where C is an arbitrary constant.*

Proof. See Appendix A.14. □

6. Numerical Illustrations

In this section, we conduct numerical studies to demonstrate the efficacy of the proposed methods. We evaluate the differences among CVaR approximation, ALSO-X, and ALSO-X+. To evaluate the effectiveness of the proposed algorithms, we use "Improvement" to denote the percentage of differences between the value of a proposed algorithm and CVaR approximation, i.e.,

$$\text{Improvement}(\%) = \frac{\text{CVaR approximation value} - \text{value of a proposed algorithm}}{|\text{CVaR approximation value}|} \times 100.$$

All the instances in this section are executed in Python 3.6 with calls to solver Gurobi (version 8.1.1 with default settings) on a personal PC with 1.6 GHz Intel Core i5 processor and 8G of memory. We set the time limit of each instance to be 3600s. Codes of the numerical experiments are available at `https://github.com/jnan97/ALSO-X`.

Now, we compare the performances of CVaR approximation, ALSO-X, and ALSO-X+ of solving the regular CCP with linear and nonlinear uncertain constraints. Particularly, we consider the number of data points $N = 400, 600, 1000$, the risk level $\varepsilon = 0.05, 0.1$, and the dimension of decision variables $n = 20, 40, 100$. For each parametric setting, we generate 5 random instances and report their average performance.

Testing a Linear CCP. Let us first consider the following linear CCP

$$v^* = \min_{\boldsymbol{x}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \boldsymbol{x} \in [0,1]^n, \frac{1}{N} \sum_{i \in [N]} \mathbb{I} \left[\sum_{j \in [n]} \xi_j^i x_j \le 100 \right] \ge 1 - \varepsilon \right\}.$$

Above, we generate the samples $\{\boldsymbol{\xi}^i\}_{i \in [N]}$ by assuming that the random parameters $\tilde{\boldsymbol{\xi}}$ are discrete and i.i.d. uniformly distributed between 1 and 50. We set $\delta_1 = 10^{-2}$ and $\delta_2 = 10^{-2}$ in ALSO-X Algorithm 1 and ALSO-X+ Algorithm 4. For each random instance, we generate the cost vector $\boldsymbol{c}$ as a random integer one with each entry uniformly distributed between -10 and -1. Since we have proven in Theorem 1 that ALSO-X delivers a better solution than that of the CVaR approximation, in ALSO-X Algorithm 1 and ALSO-X+ Algorithm 4, we use the optimal value

from CVaR approximation as an initial upper bound t_U, and the quantile bound from Ahmed et al. (2017), Song et al. (2014) as an initial lower bound t_L. Besides, at each bisection iteration, we also use the upper bound of allowable increase or lower bound of allowable decrease of the current value t to update its value for the next iteration. The numerical results for this linear CCP are displayed in Table 1. We see that although the computation time of ALSO-X is longer than that of CVaR approximation, ALSO-X Algorithm 1 can be solved within seconds and enhance the solution of CVaR approximation by around 4-8% improvement. The performance of ALSO-X+ Algorithm 4 is even more striking, which can improve the solution-quality of CVaR approximation by around 5-10%. This demonstrates the correctness and effectiveness of our proposed algorithms.

Table 1 Numerical Results for the Linear CCP

N	n	$\varepsilon = 0.05$					$\varepsilon = 0.10$				
		CVaR	ALSO-X		ALSO-X+		CVaR	ALSO-X		ALSO-X+	
		Time (s)	Time (s)	Improvement (%)	Time (s)	Improvement (%)	Time (s)	Time (s)	Improvement (%)	Time (s)	Improvement (%)
400	20	0.04	0.52	7.00	3.21	8.43	0.04	0.44	8.27	2.97	9.04
	40	0.08	0.87	4.66	6.94	6.27	0.04	0.71	6.85	7.09	7.76
	100	0.19	1.87	4.02	22.02	5.38	0.12	2.03	4.95	22.91	6.05
600	20	0.07	0.62	6.22	5.27	7.29	0.04	0.73	7.76	4.04	8.39
	40	0.10	1.29	5.36	11.43	6.28	0.04	1.09	6.56	10.72	7.46
	100	0.23	3.17	3.20	35.34	4.53	0.13	3.31	4.75	36.84	5.63
1000	20	0.12	1.07	5.86	9.24	7.10	0.06	1.04	7.45	7.75	8.33
	40	0.21	2.03	4.65	21.45	5.53	0.16	2.25	5.84	21.98	6.36
	100	0.29	4.95	4.04	60.34	4.96	0.34	5.30	5.36	60.58	6.62

Testing a Nonlinear CCP. Following Xie and Ahmed (2018b), Hong et al. (2011), Sun et al. (2014), let us consider the following chance constrained quadratic optimization problem as

$$v^* = \min_{\boldsymbol{x}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \boldsymbol{x} \in [0,1]^n, \frac{1}{N} \sum_{i\in[N]} \mathbb{I}\left[\sum_{j\in[n]} \xi_j^i x_j^2 \le 100 \right] \ge 1-\varepsilon \right\}.$$

Above, we generate the samples $\{\boldsymbol{\xi}^i\}_{i\in[N]}$ by assuming that the random parameters $\tilde{\boldsymbol{\xi}}$ are discrete and i.i.d. uniformly distributed between 1 and 99. We set $\delta_1 = 10^{-2}$ and $\delta_2 = 10^{-2}$ in ALSO-X Algorithm 1 and ALSO-X+ Algorithm 4. For each random instance, we generate its cost vector $\boldsymbol{c}$ as a random integer one with each entry uniformly distributed between -10 and -1. For this nonlinear CCP, we run SD Algorithm 2 to solve the hinge-loss approximation in Algorithm 1 as well as to solve the problem (21b), while we use Gurobi to directly solve the CVaR approximation. Note that we set the maximum number of iterations to be 50. The CVaR approximation is time-consuming, in ALSO-X Algorithm 1 and ALSO-X+ Algorithm 4, we use the greedy method to find a feasible solution as an initial upper bound t_U, and the quantile bound from Ahmed et al. (2017) as an initial lower bound t_L. The numerical results are shown in Table 2. Notably, for this nonlinear CCP, we see that both ALSO-X Algorithm 1 and ALSO-X+ Algorithm 4 provide better solutions than the CVaR approximation, and they are faster than the CVaR approximation, especially when the dimension of decision variables increases. This might be because the off-the-shelf solvers often

struggle in solving the large-scale second-order conic programs and the first-order method, on the contrary, is more effective given that the projection is relatively easy.

To demonstrate the effectiveness of our proposed method, we numerically compare the proposed ALSO-X+ Algorithm 4 with the exact Big-M method. Interested readers are referred to Appendix H for the detailed numerical results, where Big-M method is often unable to find a better solution than ALSO-X+ especially when the dimension increases and ALSO-X+ can consistently find near-optimal solutions or even optimal solutions.

Table 2 Numerical Results for the Nonlinear CCP

N	n	$\varepsilon = 0.05$					$\varepsilon = 0.10$				
		CVaR	ALSO-X		ALSO-X+		CVaR	ALSO-X		ALSO-X+	
		Time (s)	Time (s)	Improvement (%)	Time (s)	Improvement (%)	Time (s)	Time (s)	Improvement (%)	Time (s)	Improvement (%)
400	20	3.40	2.40	2.23	16.74	2.60	1.64	2.43	3.02	13.29	3.41
	40	6.90	2.61	1.99	16.84	2.81	4.58	3.02	2.27	16.53	2.72
	100	43.14	4.47	1.42	31.68	1.93	55.32	3.22	1.55	25.25	1.90
600	20	4.22	4.74	2.39	17.61	2.84	2.99	3.55	2.89	22.08	3.07
	40	11.57	4.93	1.84	18.39	2.03	8.94	5.24	2.30	22.88	2.62
	100	70.85	5.10	1.18	28.21	1.71	68.44	5.92	1.50	31.60	1.78
1000	20	7.81	5.24	2.45	20.24	2.61	7.49	7.75	2.84	34.48	2.96
	40	30.72	6.94	2.05	21.74	2.21	21.84	20.07	2.24	47.55	2.36
	100	170.28	6.89	1.30	33.50	1.54	130.08	51.41	1.59	63.30	1.73

Covering CCPs: Comparisons Between Relax-and-Scale Algorithm in Xie and Ahmed (2020) and the Proposed Algorithms. Although we have proven in Theorem 5 that for the finite-support covering CCP, the proposed ALSO-X Algorithm 1 has the same worst-case approximation ratio as the relax-and-scale algorithm (see, e.g., algorithm 2 in Ahmed and Xie 2018 or algorithm 1 in Xie and Ahmed 2020). In this subsection, we numerically compare these two algorithms as well as the CVaR approximation and ALSO-X+ Algorithm 4.

We consider the following covering CCP as

$$v^* = \min_{\boldsymbol{x}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \boldsymbol{x} \in [0,1]^n, \frac{1}{N} \sum_{i\in[N]} \mathbb{I}\left[\sum_{j\in[n]} \xi_j^i x_j \geq 40 \right] \geq 1-\varepsilon \right\}.$$

Above, we generate the samples $\{\boldsymbol{\xi}^i\}_{i\in[N]}$ by assuming that the random parameters $\tilde{\boldsymbol{\xi}}$ are discrete and i.i.d. uniformly distributed between 1 and 50. We set $\delta_1 = 10^{-2}$ and $\delta_2 = 10^{-2}$ in the ALSO-X Algorithm 1 and ALSO-X+ Algorithm 4. For each random instance, we generate the cost vector $\boldsymbol{c}$ as a random integer one with each entry uniformly distributed between 1 and 10. In the ALSO-X Algorithm 1 and ALSO-X+ Algorithm 4, the continuous relaxation bound of covering CCP (20) is set as the initial lower bound t_L, and the approximation bound $(\lfloor N\varepsilon \rfloor + 1)t_L$ is set as the initial upper bound t_U. At each bisection iteration, we also incorporate the upper bound of allowable increase or lower bound of allowable decrease of the current value of t to update its value in the next iteration. The numerical results are displayed in Table 3. We see that the proposed ALSO-X Algorithm 1 and ALSO-X+ Algorithm 4 are better than Relax-and-Scale algorithm in Ahmed and Xie 2018 in terms of solution quality, while all the three algorithms dominate the results from the CVaR approximation.

Table 3 Numerical Result for Covering CCP

N	n	$\varepsilon = 0.05$							$\varepsilon = 0.10$						
		CVaR	Relax-and-Scale Algorithm		ALSO-X		ALSO-X+		CVaR	Relax-and-Scale Algorithm		ALSO-X		ALSO-X+	
		Time (s)	Time (s)	Improvement (%)	Time (s)	Improvement (%)	Time (s)	Improvement (%)	Time (s)	Time (s)	Improvement (%)	Time (s)	Improvement (%)	Time (s)	Improvement (%)
400	20	0.02	1.17	14.57	0.43	16.45	2.11	18.78	0.01	0.63	16.45	0.42	18.57	2.21	19.85
	40	0.02	1.50	15.65	1.01	17.54	7.00	20.36	0.02	1.06	15.01	1.07	15.29	6.73	16.68
	100	0.06	2.79	7.33	2.44	10.48	18.53	12.10	0.05	2.33	9.86	2.01	11.44	15.59	13.05
600	20	0.03	1.13	17.17	0.61	18.89	4.48	20.22	0.02	1.23	18.31	0.60	19.11	4.42	19.57
	40	0.04	1.55	13.59	1.10	13.30	7.58	15.50	0.03	1.58	15.58	1.11	16.16	7.41	17.82
	100	0.09	3.26	7.83	2.88	10.80	19.22	11.55	0.07	3.00	9.20	2.43	10.77	20.88	11.69
1000	20	0.06	2.75	14.52	0.95	16.28	10.92	18.16	0.05	2.49	20.08	0.99	19.42	7.51	19.94
	40	0.12	4.01	10.40	2.28	13.47	17.65	14.53	0.09	3.99	16.75	2.15	15.82	18.98	16.68
	100	0.26	6.42	9.65	5.76	10.12	44.97	11.27	0.18	6.02	10.60	4.90	11.05	39.58	11.86

7. Conclusion

In this paper, we studied and generalized the ALSO-X algorithm for solving chance constrained programs (CCP). We showed that when uncertain constraints are convex, the ALSO-X always outperforms CVaR approximation, the well-known best convex approximation in literature. We also showed several sufficient conditions under which ALSO-X can return an optimal solution to CCP. We also provided an equivalent bilinear programming formulation of CCP, which allows us to enhance the ALSO-X with a convergent alternating minimization scheme (ALSO-X+). We extended ALSO-X to solve the distributionally robust chance constrained programs (DRCCPs) under $\infty-$Wasserstein ambiguity set. Our numerical study showed the effectiveness of the proposed algorithms.

Acknowledgment

This research has been supported by the National Science Foundation grant 2046426. Valuable comments from the associate editor and two anonymous reviewers are gratefully acknowledged.

References

Adler I, Monteiro RD (1992) A geometric view of parametric linear programming. *Algorithmica* 8(1-6):161–176.

Ahmed S, Luedtke J, Song Y, Xie W (2017) Nonanticipative duality, relaxations, and formulations for chance-constrained stochastic programs. *Mathematical Programming* 162(1-2):51–81.

Ahmed S, Papageorgiou DJ (2013) Probabilistic set covering with correlations. *Operations Research* 61(2):438–452.

Ahmed S, Shapiro A (2008) Solving chance-constrained stochastic programs via sampling and integer programming. *State-of-the-art decision-making tools in the information-intensive age*, 261–269 (Informs).

Ahmed S, Xie W (2018) Relaxations and approximations of chance constraints under finite distributions. *Mathematical Programming* 170(1):43–65.

Amaldi E, Kann V (1995) The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theoretical computer science* 147(1-2):181–210.

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust optimization* (Princeton university press).

Beraldi P, Ruszczyński A (2002) The probabilistic set-covering problem. *Operations Research* 50(6):956–967.

Bertsimas D, Shtern S, Sturt B (2018) A data-driven approach to multi-stage stochastic linear optimization. Preprint.

Bienstock D, Chertkov M, Harnett S (2014) Chance-constrained optimal power flow: Risk-aware network control under uncertainty. *Siam Review* 56(3):461–495.

Blanchet J, Murthy K (2019) Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research* 44(2):565–600.

Boyle JP, Dykstra RL (1986) A method for finding projections onto the intersection of convex sets in hilbert spaces. *Advances in order restricted statistical inference*, 28–47 (Springer).

Calafiore GC, Campi MC (2006) The scenario approach to robust control design. *IEEE Transactions on automatic control* 51(5):742–753.

Charnes A, Cooper WW (1963) Deterministic equivalents for optimizing and satisficing under chance constraints. *Operations research* 11(1):18–39.

Charnes A, Cooper WW, Symonds GH (1958) Cost horizons and certainty equivalents: an approach to stochastic programming of heating oil. *Management science* 4(3):235–263.

Chen Z, Kuhn D, Wiesemann W (2018) Data-driven chance constrained programs over wasserstein balls. arXiv preprint arXiv:1809.00210.

Chen Z, Xie W (2019) Sharing the value-at-risk under distributional ambiguity. Available at SSRN 3400033.

Chen Z, Xie W (2020) Regret in the newsvendor model with demand and yield randomness. Available at SSRN.

Chow G (1995) Portfolio selection based on return, risk, and relative performance. *Financial Analysts Journal* 51(2):54–60.

Deng Y, Shen S (2016) Decomposition algorithms for optimizing multi-server appointment scheduling with chance constraints. *Mathematical Programming* 157(1):245–276.

Dentcheva D, Prékopa A, Ruszczynski A (2000) Concavity and efficient points of discrete distributions in probabilistic programming. *Mathematical Programming* 89(1):55–77.

Embrechts P, McNeil A, Straumann D (2002) Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond* 1:176–223.

Fabozzi FJ, Markowitz HM, Kolm PN, Gupta F (2012) Mean-variance model for portfolio selection. *Encyclopedia of Financial Models* .

Gao R, Kleywegt AJ (2016) Distributionally robust stochastic optimization with wasserstein distance. arXiv preprint arXiv:1604.02199.

Garey MR (1979) A guide to the theory of np-completeness.

Hanasusanto GA, Roitch V, Kuhn D, Wiesemann W (2015) A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Mathematical Programming* 151(1):35–62.

Hanasusanto GA, Roitch V, Kuhn D, Wiesemann W (2017) Ambiguous joint chance constraints under mean and dispersion information. *Operations Research* 65(3):751–767.

Henrion R (2006) Some remarks on value-at-risk optimization. *International Journal of Management Science and Engineering Management* 1(2):111–118.

Henrion R (2007) Structural properties of linear probabilistic constraints. *Optimization* 56(4):425–440.

Henrion R, Strugarek C (2008) Convexity of chance constraints with independent random variables. *Computational Optimization and Applications* 41(2):263–276.

Henrion R, Strugarek C (2011) Convexity of chance constraints with dependent random variables: the use of copulae. *Stochastic optimization methods in finance and energy*, 427–439 (Springer).

Hong LJ, Yang Y, Zhang L (2011) Sequential convex approximations to joint chance constrained programs: A monte carlo approach. *Operations Research* 59(3):617–630.

Jaworski P, Durante F, Hardle WK, Rychlik T (2010) *Copula theory and its applications*, volume 198 (Springer).

Ji R, Lejeune M (2020) Data-driven distributionally robust chance-constrained optimization with wasserstein metric. Available at SSRN 3201356.

Kall P, Wallace SW, Kall P (1994) *Stochastic programming* (Springer).

Kamdem JS (2005) Value-at-risk and expected shortfall for linear portfolios with elliptically distributed risk factors. *International Journal of Theoretical and Applied Finance* 8(05):537–551.

Kataoka S (1963) A stochastic programming model. *Econometrica: Journal of the Econometric Society* 181–196.

Lagoa CM, Li X, Sznaier M (2005) Probabilistically constrained linear programs and risk-adjusted controller design. *SIAM Journal on Optimization* 15(3):938–951.

Landsman ZM, Valdez EA (2003) Tail conditional expectations for elliptical distributions. *North American Actuarial Journal* 7(4):55–71.

Lejeune MA, Margot F (2016) Solving chance-constrained optimization problems with stochastic quadratic inequalities. *Operations Research* 64(4):939–957.

Luedtke J, Ahmed S (2008) A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization* 19(2):674–699.

Luedtke J, Ahmed S, Nemhauser GL (2010) An integer programming approach for linear programs with probabilistic constraints. *Mathematical programming* 122(2):247–272.

Markowitz HM (1991) Foundations of portfolio theory. *The journal of finance* 46(2):469–477.

Nemirovski A (2001) Lectures on modern convex optimization. *Society for Industrial and Applied Mathematics (SIAM* (Citeseer).

Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programmings. *SIAM Journal on Optimization* 19(4):1574–1609.

Nemirovski A, Shapiro A (2006) Scenario approximations of chance constraints. *Probabilistic and randomized methods for design under uncertainty*, 3–47 (Springer).

Nemirovski A, Shapiro A (2007) Convex approximations of chance constrained programs. *SIAM Journal on Optimization* 17(4):969–996.

Nesterov Y (2003) *Introductory lectures on convex optimization: A basic course*, volume 87 (Springer Science & Business Media).

Pagnoncelli BK, Ahmed S, Shapiro A (2009) Sample average approximation method for chance constrained programming: theory and applications. *Journal of optimization theory and applications* 142(2):399–416.

Pena-Ordieres A, Luedtke JR, Wachter A (2020) Solving chance-constrained problems via a smooth sample-based nonlinear approximation. *SIAM Journal on Optimization* 30(3):2221–2250.

Prékopa A (1974) Programming under probabilistic constraints with a random technology matrix. *Statistics: A Journal of Theoretical and Applied Statistics* 5(2):109–116.

Prékopa A (2013) *Stochastic programming*, volume 324 (Springer Science & Business Media).

Qiu F, Ahmed S, Dey SS, Wolsey LA (2014) Covering linear programming with violations. *INFORMS Journal on Computing* 26(3):531–546.

Rahimian H, Mehrotra S (2019) Distributionally robust optimization: A review. arXiv preprint arXiv:1908.05659.

Rockafellar RT, Wets RJ (1982) On the interchange of subdifferentiation and conditional expectation for convex functionals. *Stochastics: An International Journal of Probability and Stochastic Processes* 7(3):173–182.

Royden HL, Fitzpatrick P (1988) *Real analysis*, volume 32 (Macmillan New York).

Rudin W, et al. (1964) *Principles of mathematical analysis*, volume 3 (McGraw-hill New York).

Ruszczyński A (2002) Probabilistic programming with discrete distributions and precedence constrained knapsack polyhedra. *Mathematical Programming* 93(2):195–215.

Shiina T (1999) Numerical solution technique for joint chance-constrained programming problem: An application to electric power capacity expansion. *Journal of the Operations Research Society of Japan* 42(2):128–140.

Slater M (2014) Lagrange multipliers revisited. *Traces and emergence of nonlinear programming*, 293–306 (Springer).

Song Y, Luedtke JR, Küçükyavuz S (2014) Chance-constrained binary packing problems. *INFORMS Journal on Computing* 26(4):735–747.

Sun H, Xu H, Wang Y (2014) Asymptotic analysis of sample average approximation for stochastic optimization problems with joint chance constraints via conditional value at risk and difference of convex functions. *Journal of Optimization Theory and Applications* 161(1):257–284.

Suykens JA, Vandewalle J (1999) Least squares support vector machine classifiers. *Neural processing letters* 9(3):293–300.

Takyi AK, Lence BJ (1999) Surface water quality management using a multiple-realization chance constraint method. *Water Resources Research* 35(5):1657–1670.

Talluri S, Narasimhan R, Nair A (2006) Vendor performance with supply risk: A chance-constrained dea approach. *International Journal of Production Economics* 100(2):212–222.

Tao PD, An LTH (1997) Convex analysis approach to dc programming: theory, algorithms and applications. *Acta mathematica vietnamica* 22(1):289–355.

Tavakoli R (2016) On the coupled continuous knapsack problems: projection onto the volume constrained gibbs n-simplex. *Optimization Letters* 10(1):137–158.

Terkelsen F (1972) Some minimax theorems. *Mathematica Scandinavica* 31(2):405–413.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1):267–288.

Tijms HC (2003) *A first course in stochastic models* (John Wiley and sons).

Van Ackooij W, Malick J (2019) Eventual convexity of probability constraints with elliptical distributions. *Mathematical Programming* 175(1):1–27.

Xie W (2019) On distributionally robust chance constrained programs with wasserstein distance. *Mathematical Programming* 1–41.

Xie W (2020) Tractable reformulations of two-stage distributionally robust linear programs over the type-∞ wasserstein ball. *Operations Research Letters* 48(4):513–523.

Xie W, Ahmed S (2017) Distributionally robust chance constrained optimal power flow with renewables: A conic reformulation. *IEEE Transactions on Power Systems* 33(2):1860–1867.

Xie W, Ahmed S (2018a) On deterministic reformulations of distributionally robust joint chance constrained optimization problems. *SIAM Journal on Optimization* 28(2):1151–1182.

Xie W, Ahmed S (2018b) On quantile cuts and their closure for chance constrained optimization problems. *Mathematical Programming* 172(1-2):621–646.

Xie W, Ahmed S (2020) Bicriteria approximation of chance-constrained covering problems. *Operations Research* 68(2):516–533.

Zhang J, Xie W, Sarin SC (2021) Robust multi-product newsvendor model with uncertain demand and substitution. *European Journal of Operational Research* 293(1):190–202.

Zhang Y, Shen S, Mathieu JL (2016) Distributionally robust chance-constrained optimal power flow with uncertain renewables and uncertain reserves provided by loads. *IEEE Transactions on Power Systems* 32(2):1378–1388.

Zymler S, Kuhn D, Rustem B (2013) Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming* 137(1-2):167–198.

Appendix A. Proofs

Proofs in Section 2

A.1 Proof of Proposition 1

Proposition 1 *The CCP* (1) *can be viewed as the following equivalent form*

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}, z(\cdot), s(\cdot)} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \le s(\tilde{\boldsymbol{\xi}}), \mathbb{E}[z(\tilde{\boldsymbol{\xi}})] \ge 1-\varepsilon, z(\tilde{\boldsymbol{\xi}}) \in [0,1], \mathbb{E}[z(\tilde{\boldsymbol{\xi}}) s(\tilde{\boldsymbol{\xi}})] = 0, s(\tilde{\boldsymbol{\xi}}) \ge 0 \right\}. \quad (3)$$

Proof. We prove Formulation (3) and CCP (2) are equivalent. Let v_1, v_2 be the optimal values of Formulation (3) and CCP (2), respectively. Then it remains to show that $v_1 \le v_2$ and $v_2 \le v_1$.

1. ($v_1 \le v_2$) Let $(\boldsymbol{x}^*, z^*(\cdot))$ be an optimal solution to CCP (2). Define $s^*(\boldsymbol{\xi}) := \max\{g(\boldsymbol{x}^*, \boldsymbol{\xi}), 0\}$. According to the properties of the measurable functions (see, e.g., section 3.1 in Royden and Fitzpatrick 1988), $s^*(\cdot)$ is measurable. As in Formulation (2), the constraint $z^*(\tilde{\boldsymbol{\xi}}) \le \mathbb{I}(g(\boldsymbol{x}^*, \tilde{\boldsymbol{\xi}}) \le 0)$ holds a.s., we have

$$0 \le \mathbb{E}[z^*(\tilde{\boldsymbol{\xi}}) s^*(\tilde{\boldsymbol{\xi}})] \le \mathbb{E}\left[\mathbb{I}(g(\boldsymbol{x}^*, \tilde{\boldsymbol{\xi}}) \le 0) s^*(\tilde{\boldsymbol{\xi}}) \right],$$

where the first inequality is due to the nonnegativity of $z^*(\boldsymbol{\xi}) s^*(\boldsymbol{\xi})$ and the second one is because of monotonicity and nonnegativity of $s^*(\cdot)$. Since $s^*(\boldsymbol{\xi}) := \max\{g(\boldsymbol{x}^*, \boldsymbol{\xi}), 0\}$, for any positive $t > 0$, we have

$$\mathbb{P}\left\{ \tilde{\boldsymbol{\xi}} : \mathbb{I}(g(\boldsymbol{x}^*, \tilde{\boldsymbol{\xi}}) \le 0) s^*(\tilde{\boldsymbol{\xi}}) \ge t \right\} = \mathbb{P}\left\{ \tilde{\boldsymbol{\xi}} : s^*(\tilde{\boldsymbol{\xi}}) \ge t, g(\boldsymbol{x}^*, \tilde{\boldsymbol{\xi}}) \le 0 \right\} = 0,$$

which implies

$$\mathbb{E}\left[\mathbb{I}(g(\boldsymbol{x}^*, \tilde{\boldsymbol{\xi}}) \le 0) s^*(\tilde{\boldsymbol{\xi}}) \right] \le 0.$$

Thus, we must have $\mathbb{E}[z^*(\tilde{\boldsymbol{\xi}}) s^*(\tilde{\boldsymbol{\xi}})] = 0$. Therefore, $(\boldsymbol{x}^*, z^*(\cdot), s^*(\cdot))$ is feasible to Formulation (3), and hence $v_1 \le v_2$.

2. ($v_2 \le v_1$) Let $(\boldsymbol{x}^*, z^*(\cdot), s^*(\cdot))$ be an optimal solution to Formulation (3) and suppose $\widehat{z}^*(\cdot) = \mathbb{I}\{z^*(\cdot) > 0\}$. By the properties of the measurable functions (see, e.g., section 3.1 in Royden and Fitzpatrick 1988), the functional variable $\widehat{z}^*(\cdot)$ is measurable. Thus, $\mathbb{P}\{\tilde{\boldsymbol{\xi}} \colon \widehat{z}^*(\tilde{\boldsymbol{\xi}}) \in \{0,1\}\} = 1$ and $\mathbb{P}\{\tilde{\boldsymbol{\xi}} \colon \widehat{z}^*(\tilde{\boldsymbol{\xi}}) \ge z^*(\tilde{\boldsymbol{\xi}})\} = 1$. Together with the fact that $\mathbb{E}[z^*(\tilde{\boldsymbol{\xi}})] \ge 1-\varepsilon$, we have $\mathbb{E}[\widehat{z}^*(\tilde{\boldsymbol{\xi}})] \ge 1-\varepsilon$ because of monotonicity. According to constraints $\mathbb{E}[z^*(\tilde{\boldsymbol{\xi}}) s^*(\tilde{\boldsymbol{\xi}})] = 0, z^*(\tilde{\boldsymbol{\xi}}) \in [0,1], s^*(\tilde{\boldsymbol{\xi}}) \ge 0$ almost surely, we have

$$\mathbb{P}\left\{ \tilde{\boldsymbol{\xi}} \colon z^*(\tilde{\boldsymbol{\xi}}) s^*(\tilde{\boldsymbol{\xi}}) = 0, 0 \le z^*(\tilde{\boldsymbol{\xi}}) \le 1, s^*(\tilde{\boldsymbol{\xi}}) \ge 0 \right\} = 1.$$

By the law of total probability (see, e.g., appendix A of Tijms 2003), the above identity is equivalent to

$$\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon 0\le z^*(\tilde{\boldsymbol{\xi}})\le 1, s^*(\tilde{\boldsymbol{\xi}})=0\right\}+\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon z^*(\tilde{\boldsymbol{\xi}})=0, s^*(\tilde{\boldsymbol{\xi}})>0\right\}=1. \tag{33a}$$

Next, we bound two terms on the left-hand side separately. Since $g(\boldsymbol{x}^*,\tilde{\boldsymbol{\xi}})\le s^*(\tilde{\boldsymbol{\xi}})$ holds almost surely, the conditional probability $\mathbb{P}\{\tilde{\boldsymbol{\xi}}\colon \mathbb{I}\{z^*(\tilde{\boldsymbol{\xi}})>0\}\le\mathbb{I}\{g(\boldsymbol{x}^*,\tilde{\boldsymbol{\xi}})\le 0\}\mid 0\le z^*(\tilde{\boldsymbol{\xi}})\le 1, s^*(\tilde{\boldsymbol{\xi}})=0\}=1$. Hence, the first term on the left-hand side in (33a) is equivalent to

$$\begin{aligned}&\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon 0\le z^*(\tilde{\boldsymbol{\xi}})\le 1, s^*(\tilde{\boldsymbol{\xi}})=0\right\}\\ =&\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon \mathbb{I}\left\{z^*(\tilde{\boldsymbol{\xi}})>0\right\}\le\mathbb{I}\left\{g(\boldsymbol{x}^*,\tilde{\boldsymbol{\xi}})\le 0\right\}, 0\le z^*(\tilde{\boldsymbol{\xi}})\le 1, s^*(\tilde{\boldsymbol{\xi}})=0\right\}.\end{aligned}$$

Since $\widehat{z}^*(\tilde{\boldsymbol{\xi}})=\mathbb{I}(z^*(\tilde{\boldsymbol{\xi}})>0)$ and $0\le z^*(\tilde{\boldsymbol{\xi}})\le 1$ hold almost surely, we have

$$\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon 0\le z^*(\tilde{\boldsymbol{\xi}})\le 1, s^*(\tilde{\boldsymbol{\xi}})=0\right\}=\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon \widehat{z}^*(\tilde{\boldsymbol{\xi}})\le\mathbb{I}\left\{g(\boldsymbol{x}^*,\tilde{\boldsymbol{\xi}})\le 0\right\}, s^*(\tilde{\boldsymbol{\xi}})=0\right\}. \tag{33b}$$

Similarly, the second term on the left-hand side in (33a) can be written as

$$\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon z^*(\tilde{\boldsymbol{\xi}})=0, s^*(\tilde{\boldsymbol{\xi}})>0\right\}=\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon \widehat{z}^*(\tilde{\boldsymbol{\xi}})\le\mathbb{I}\left\{g(\boldsymbol{x}^*,\tilde{\boldsymbol{\xi}})\le 0\right\}, s^*(\tilde{\boldsymbol{\xi}})>0\right\}. \tag{33c}$$

Combining equalities (33a), (33b) and (33c) together, we have

$$\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon \widehat{z}^*(\tilde{\boldsymbol{\xi}})\le\mathbb{I}\left\{g(\boldsymbol{x}^*,\tilde{\boldsymbol{\xi}})\le 0\right\}, s^*(\tilde{\boldsymbol{\xi}})=0\right\}+\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon \widehat{z}^*(\tilde{\boldsymbol{\xi}})\le\mathbb{I}\left\{g(\boldsymbol{x}^*,\tilde{\boldsymbol{\xi}})\le 0\right\}, s^*(\tilde{\boldsymbol{\xi}})>0\right\}=1.$$

By the law of total probability (see, e.g., appendix A of Tijms 2003), the above equality can be simplified as

$$\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon \widehat{z}^*(\tilde{\boldsymbol{\xi}})\le\mathbb{I}\left\{g(\boldsymbol{x}^*,\tilde{\boldsymbol{\xi}})\le 0\right\}\right\}=1.$$

Thus, $(\boldsymbol{x}^*,\widehat{z}^*(\cdot))$ satisfies the constraints in CCP (2). Therefore, $(\boldsymbol{x}^*,\widehat{z}^*(\cdot))$ is feasible to CCP (2), and thus $v_2\le v_1$.

Therefore, Formulation (3) and CCP (2) are equivalent. This concludes the proof. □

A.2 Proof of Proposition 2

Proposition 2 *CCP* (4) *is equivalent to*

$$v^*=\min_t\ t, \tag{5a}$$

$$s.t.\ (\boldsymbol{x}^*,s^*(\cdot),z^*(\cdot))\in\operatorname*{arg\,min}_{\substack{\boldsymbol{x}\in\mathcal{X},\\ z(\cdot)\in[0,1],\\ s(\cdot)\ge 0}}\left\{\mathbb{E}[z(\tilde{\boldsymbol{\xi}})s(\tilde{\boldsymbol{\xi}})]\colon g(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\le s(\tilde{\boldsymbol{\xi}}),\mathbb{E}[z(\tilde{\boldsymbol{\xi}})]\ge 1-\varepsilon,\boldsymbol{c}^\top\boldsymbol{x}\le t\right\}, \tag{5b}$$

$$\mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon g(\boldsymbol{x}^*,\tilde{\boldsymbol{\xi}})\le 0\right\}\ge 1-\varepsilon. \tag{5c}$$

Proof. Let v_1, v_2 be the optimal values of CCP (3) and Formulation (5), respectively. Then it remains to show that $v_1 \le v_2$ and $v_2 \le v_1$.

($v_1 \le v_2$) Let $(\boldsymbol{x}^*, z^*(\cdot), s^*(\cdot), t^*)$ be an optimal solution of Formulation (5). According to (5c), we have $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : g(\boldsymbol{x}^*, \tilde{\boldsymbol{\xi}}) \le 0\} \ge 1-\varepsilon$, which implies $\mathbb{E}_{\mathbb{P}}[\mathbb{I}(g(\boldsymbol{x}^*, \tilde{\boldsymbol{\xi}}) \le 0)] \ge 1-\varepsilon$. We can construct an optimal solution of Formulation (5) as follows. Let $\widehat{s}(\boldsymbol{\xi}) = \max\{g(\boldsymbol{x}^*, \boldsymbol{\xi}), 0\}$ and $\widehat{z}(\boldsymbol{\xi}) = \mathbb{I}\{g(\boldsymbol{x}^*, \boldsymbol{\xi}) \le 0\}$. According to the properties of the measurable functions (see, e.g., section 3.1 in Royden and Fitzpatrick 1988), $\widehat{s}(\tilde{\boldsymbol{\xi}})$ and $\widehat{z}(\tilde{\boldsymbol{\xi}})$ are measurable. In this way, $(\boldsymbol{x}^*, \widehat{z}(\boldsymbol{\xi}), \widehat{s}(\boldsymbol{\xi}))$ satisfies the constraints in (5b). Since $\widehat{s}(\tilde{\boldsymbol{\xi}}) \ge 0$, $\widehat{z}(\tilde{\boldsymbol{\xi}}) \ge 0$ hold almost surely, $\mathbb{E}[\widehat{z}(\tilde{\boldsymbol{\xi}})\widehat{s}(\tilde{\boldsymbol{\xi}})]$ is well defined. From the proof in Proposition 1, we have $\mathbb{E}[\widehat{z}(\tilde{\boldsymbol{\xi}})\widehat{s}(\tilde{\boldsymbol{\xi}})] = 0$, indicating $(\boldsymbol{x}^*, \widehat{z}(\cdot), \widehat{s}(\cdot))$ solves the lower-level problem (5b). Thus, $(\boldsymbol{x}^*, \widehat{z}(\cdot), \widehat{s}(\cdot), t^*)$ is another optimal solution of Formulation (5). The fact that $(\boldsymbol{x}^*, \widehat{z}(\cdot), \widehat{s}(\cdot))$ is feasible to CCP (3) implies $v_1 \le v_2$.

($v_2 \le v_1$) Let $(\boldsymbol{x}^*, z^*(\cdot), s^*(\cdot))$ be an optimal solution to CCP (3) and $t^* = \boldsymbol{c}^\top \boldsymbol{x}^*$. We have $\mathbb{E}[z^*(\tilde{\boldsymbol{\xi}})s^*(\tilde{\boldsymbol{\xi}})] = 0$, which solves the lower-level problem (5b). Therefore, $(\boldsymbol{x}^*, z^*(\cdot), s^*(\cdot), t^*)$ satisfies the constraints in Formulation (5). Thus, $v_2 \le v_1$. □

A.3 Proof of Proposition 4

Proposition 4 *For any elliptical distribution* $\mathbb{P}_{\mathrm{E}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{g})$*,* ALSO-X (7) *corresponding to the single linear CCP admits the following form*

$$v^A = \min_{t} \quad t, \tag{10a}$$

$$s.t. \quad (\boldsymbol{x}^*, \alpha^*) \in \arg\min_{\boldsymbol{x}\in\mathcal{X},\alpha} \left\{ \sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x})} \left(\overline{G}(\alpha^2/2) - \alpha + \alpha\Phi(\alpha)\right) : \boldsymbol{c}^\top \boldsymbol{x} \le t, \frac{b_1(\boldsymbol{x}) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x})}{\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x})}} = \alpha \right\}, \tag{10b}$$

$$b_1(\boldsymbol{x}^*) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x}^*) \ge \Phi^{-1}(1-\varepsilon)\sqrt{\boldsymbol{a}_1(\boldsymbol{x}^*)^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x}^*)}, \tag{10c}$$

where $\overline{G}(\tau) = G(\infty) - G(\tau)$ *and* $G(\tau) = k\int_0^\tau \widehat{g}(z)dz$*. By default, we let* $\frac{0}{0} = 0$ *and* $\frac{c}{0} = \mathrm{sign}(c)\infty$ *if* $c \ne 0$*.*

Proof. It is sufficient to prove that the hinge-loss approximation (6) is equivalent to (10b). In fact, the objective function in the hinge-loss approximation (6) can be calculated based on the definition of conditional expectation

$$\mathbb{E}\left[[\tilde{\boldsymbol{\xi}}^\top \boldsymbol{a}_1(\boldsymbol{x}) - b_1(\boldsymbol{x})]_+\right] = \mathbb{E}\left[\tilde{\boldsymbol{\xi}}^\top \boldsymbol{a}_1(\boldsymbol{x}) - b_1(\boldsymbol{x}) \mid \tilde{\boldsymbol{\xi}}^\top \boldsymbol{a}_1(\boldsymbol{x}) - b_1(\boldsymbol{x}) \ge 0\right] \mathbb{P}\left\{\tilde{\boldsymbol{\xi}}^\top \boldsymbol{a}_1(\boldsymbol{x}) - b_1(\boldsymbol{x}) \ge 0\right\}.$$

Using the cumulative distribution function formula of an elliptical distribution, the objective function can be further simplified as

$$\left(1 - \Phi\left(\frac{b_1(\boldsymbol{x}) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x})}{\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x})}}\right)\right) \mathbb{E}\left[\tilde{\boldsymbol{\xi}}^\top \boldsymbol{a}_1(\boldsymbol{x}) - b_1(\boldsymbol{x}) \mid \tilde{\boldsymbol{\xi}}^\top \boldsymbol{a}_1(\boldsymbol{x}) - b_1(\boldsymbol{x}) \ge 0\right].$$

According to the closed-form expression of the expectation of the truncated elliptical distribution (see, e.g., theorem 1 in Landsman and Valdez 2003), the objective function is equivalent to

$$\left(1-\Phi(\frac{b_1(\boldsymbol{x})-\boldsymbol{\mu}^\top\boldsymbol{x}}{\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}})\right)\left(\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x})-b_1(\boldsymbol{x})+\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}\frac{\overline{G}(\frac{1}{2}(\frac{b_1(\boldsymbol{x})-\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x})}{\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}})^2)}{\left(1-\Phi(\frac{b_1(\boldsymbol{x})-\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x})}{\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}})\right)}\right).$$

Let $\alpha=(b_1(\boldsymbol{x})-\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x}))/\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}$. Then the objective can be simplified as

$$(1-\Phi(\alpha))\left(\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x})-b_1(\boldsymbol{x})\right)+\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}\overline{G}(\alpha^2/2)=\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}\left(\overline{G}(\alpha^2/2)-\alpha+\alpha\Phi(\alpha)\right).$$

This concludes the proof. □

Proofs in Section 3

A.4 Proof of Theorem 1

Theorem 1 *Let v^A, v^{CVaR} denote the optimal value of the* ALSO-X (7) *and the* CVaR *approximation* (13)*, respectively. Then, under Assumptions A1-A2, we must have $v^A\le v^{\mathrm{CVaR}}$.*

Proof. It is sufficient to show that for any given t, if an optimal solution of the lower-level problem (7b) violates the chance constraint (7c), so does CVaR approximation. Next, we split the proof into two steps.

Step 1. Recall that for any given t, the hinge-loss approximation (7b) is

$$\begin{aligned} v^A(t)=\min_{\boldsymbol{x}\in\mathcal{X},s(\cdot)}\quad & \mathbb{E}[s(\tilde{\boldsymbol{\xi}})], && (34a)\\ \text{s.t.}\quad & g(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\le s(\tilde{\boldsymbol{\xi}}), && (34b)\\ & \boldsymbol{c}^\top\boldsymbol{x}\le t, && (34c)\\ & s(\tilde{\boldsymbol{\xi}})\ge 0. && (34d)\end{aligned}$$

And the lower-level problem (13b) is

$$\begin{aligned} v^{\mathrm{CVaR}}(t)=\min_{\boldsymbol{x}\in\mathcal{X},s(\cdot),\beta\le 0}\quad & \mathbb{E}[s(\tilde{\boldsymbol{\xi}})]-(1-\varepsilon)\beta, && (35a)\\ \text{s.t.}\quad & g(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\le s(\tilde{\boldsymbol{\xi}}), && (35b)\\ & \boldsymbol{c}^\top\boldsymbol{x}\le t, && (35c)\\ & s(\tilde{\boldsymbol{\xi}})\ge \beta. && (35d)\end{aligned}$$

Suppose that $(\boldsymbol{x}^*,s^*(\cdot))$ is an optimal solution to the hinge-loss approximation problem (34). We would like to prove that if $\boldsymbol{x}^*$ is infeasible to CCP (1), i.e., $\mathbb{P}\{\tilde{\boldsymbol{\xi}}\colon s^*(\tilde{\boldsymbol{\xi}})>0\}>\varepsilon$, then we must have

$v^A(t) = v^{\mathrm{CVaR}}(t) > 0$. Note that we already have $v^A(t) > 0$ due the presumption that $\boldsymbol{x}^*$ is infeasible to CCP (1). To prove $v^A(t) = v^{\mathrm{CVaR}}(t)$, we let $\alpha(\cdot), \pi, \mu(\cdot)$ be the dual variables of constraints (35b), (35c), and (35d), respectively. The Lagrangian of the lower-level problem (35) is

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}, \beta, s(\cdot), \mu(\cdot), \alpha(\cdot), \pi) :=& \mathbb{E}[s(\tilde{\boldsymbol{\xi}})] - (1-\varepsilon)\beta + \mathbb{E}\left[\mu(\tilde{\boldsymbol{\xi}})^\top [\beta - s(\tilde{\boldsymbol{\xi}})]\right] \\ &+ \pi(\boldsymbol{c}^\top \boldsymbol{x} - t) + \mathbb{E}\left[\alpha(\tilde{\boldsymbol{\xi}})^\top [g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - s(\tilde{\boldsymbol{\xi}})]\right],\end{aligned}$$

and its dual problem can be written as

$$v_D^{\mathrm{CVaR}}(t) = \max_{\mu(\cdot), \alpha(\cdot), \pi} \min_{\boldsymbol{x}, \beta, s(\cdot)} \mathcal{L}(\boldsymbol{x}, \beta, s(\cdot), \mu(\cdot), \alpha(\cdot), \pi). \tag{36a}$$

According to Assumptions A1-A2, the relaxed Slater condition holds, and thus theorem 1 in Slater (2014) implies that there is no duality gap between the lower-level problem (35) and its dual. Let $(\widehat{\boldsymbol{x}}, \widehat{s}(\cdot), \widehat{\beta})$ be an optimal solution of (35) and $(\widehat{\alpha}(\cdot), \widehat{\mu}(\cdot), \widehat{\pi})$ be an optimal solution of (36a). Then we have the following necessary and sufficient KKT conditions:

$$\begin{aligned}&-\widehat{\pi}\boldsymbol{c}^\top \in \partial_x \mathbb{E}\left[\widehat{\alpha}(\tilde{\boldsymbol{\xi}})^\top [g(\widehat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})]\right] + \mathcal{N}_X(\widehat{\boldsymbol{x}}), \mathbb{E}[\widehat{\mu}(\tilde{\boldsymbol{\xi}})] \le 1-\varepsilon, \widehat{\mu}(\tilde{\boldsymbol{\xi}}) + \widehat{\alpha}(\tilde{\boldsymbol{\xi}}) = 1, \\ &0 \le \widehat{\alpha}(\tilde{\boldsymbol{\xi}}) \perp \left(\widehat{s}(\tilde{\boldsymbol{\xi}}) - g(\widehat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})\right) \ge 0, 0 \le \widehat{\pi} \perp (t - \boldsymbol{c}^\top \widehat{\boldsymbol{x}}) \ge 0, 0 \le \widehat{\mu}(\tilde{\boldsymbol{\xi}}) \perp \widehat{s}(\tilde{\boldsymbol{\xi}}) - \widehat{\beta} \ge 0, \\ &\widehat{\beta} \le 0, \widehat{\boldsymbol{x}} \in \mathcal{X}.\end{aligned} \tag{KKT1}$$

Similarly, let $\alpha(\cdot), \pi, \mu(\cdot)$ be the dual variables of constraints (34b), (34c), and (34d), respectively. The Lagrangian of the hinge-loss approximation (34) is

$$\mathcal{L}(\boldsymbol{x}, s(\cdot), \mu(\cdot), \alpha(\cdot), \pi) := \mathbb{E}[s(\tilde{\boldsymbol{\xi}})] + \mathbb{E}\left[\mu(\tilde{\boldsymbol{\xi}})^\top [-s(\tilde{\boldsymbol{\xi}})]\right] + \pi(\boldsymbol{c}^\top \boldsymbol{x} - t) + \mathbb{E}\left[\alpha(\tilde{\boldsymbol{\xi}})^\top [g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) - s(\tilde{\boldsymbol{\xi}})]\right],$$

and its dual program is

$$v_D^A(t) = \max_{\mu(\cdot), \alpha(\cdot), \pi} \min_{\boldsymbol{x}, s(\cdot)} \mathcal{L}(\boldsymbol{x}, s(\boldsymbol{\xi}), \mu(\boldsymbol{\xi}), \alpha(\boldsymbol{\xi}), \pi). \tag{36b}$$

From the similar argument, the strong duality also holds, i.e., we must have $v_D^A(t) = v^A(t)$. Let $(\bar{\boldsymbol{x}}, \bar{s}(\cdot))$ be an optimal solution of (34), and $(\bar{\alpha}(\cdot), \bar{\mu}(\cdot), \bar{\pi})$ be an optimal dual solution of (36b). Then we have the following necessary and sufficient KKT conditions:

$$\begin{aligned}&-\bar{\pi}\boldsymbol{c}^\top \in \partial_x \mathbb{E}\left[\bar{\alpha}(\tilde{\boldsymbol{\xi}})^\top [g(\bar{x}, \tilde{\boldsymbol{\xi}})]\right] + \mathcal{N}_X(\bar{\boldsymbol{x}}), \bar{\mu}(\tilde{\boldsymbol{\xi}}) + \bar{\alpha}(\tilde{\boldsymbol{\xi}}) = 1, \bar{\boldsymbol{x}} \in \mathcal{X}, \\ &0 \le \bar{\alpha}(\tilde{\boldsymbol{\xi}}) \perp \left(\bar{s}(\tilde{\boldsymbol{\xi}}) - g(\bar{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})\right) \ge 0, 0 \le \bar{\pi} \perp (t - \boldsymbol{c}^\top \bar{\boldsymbol{x}}) \ge 0, 0 \le \bar{\mu}(\tilde{\boldsymbol{\xi}}) \perp \bar{s}(\tilde{\boldsymbol{\xi}}) \ge 0.\end{aligned} \tag{KKT2}$$

Step 2. To prove $v^{\mathrm{CVaR}}(t) = v^A(t)$ is equivalent to show that $v_D^{\mathrm{CVaR}}(t) = v_D^A(t)$. According to our presumption that $(\bar{\boldsymbol{x}}, \bar{s}(\cdot))$ violates the chance constraint (7c), i.e., $\mathbb{P}\{\tilde{\boldsymbol{\xi}} \colon \bar{s}(\tilde{\boldsymbol{\xi}}) > 0\} > \varepsilon$, which implies that $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : \bar{\mu}(\tilde{\boldsymbol{\xi}}) = 0\} > \varepsilon$ from conditions (KKT2). We also have

$$\mathbb{E}\left[\bar{\mu}(\tilde{\boldsymbol{\xi}}) + \bar{\alpha}(\tilde{\boldsymbol{\xi}})\right] = 1,$$

and

$$\mathbb{E}\left[\bar{\alpha}(\tilde{\boldsymbol{\xi}})\right] \geq \mathbb{E}\left[\bar{\alpha}(\tilde{\boldsymbol{\xi}})\mathbb{I}(\bar{\mu}(\tilde{\boldsymbol{\xi}})=0)\right] = \mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon \bar{\mu}(\tilde{\boldsymbol{\xi}})=0\right\} > \varepsilon.$$

Since $\bar{\alpha}(\cdot) \geq 0$, thus $\mathbb{E}[\bar{\mu}(\tilde{\boldsymbol{\xi}})] < 1-\varepsilon$ must hold. This implies that the primal and dual pair, $(\bar{\boldsymbol{x}}, \bar{s}(\cdot), \bar{\beta}=0)$ and $(\bar{\alpha}(\cdot), \bar{\pi}, \bar{\mu}(\cdot))$, satisfies conditions (KKT1). That is, $(\bar{\boldsymbol{x}}, \bar{s}(\cdot), \bar{\beta}=0)$ is optimal to the lower-level problem (36a). Hence, we have $v^A(t) = v^{\mathrm{CVaR}}(t) > 0$. □

A.5 Proof of Theorem 2

Theorem 2 *For Special Case 1, the following results must hold:*

(i) For any feasible pair of $(t, a(\cdot))$ under which the hinge-loss approximation (16b) *has a feasible solution $(\boldsymbol{x}, s(\cdot))$ satisfying $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : s(\tilde{\boldsymbol{\xi}})=0\} \geq 1-\varepsilon$, every optimal solution $(\boldsymbol{x}^*, s^*(\cdot))$ to the hinge-loss approximation* (16b) *shares the same $s^*(\cdot)$ and satisfies $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : s^*(\tilde{\boldsymbol{\xi}})=0\} \geq 1-\varepsilon$, if and only if the generalized nullspace property holds; and*

(ii) Suppose that the generalized nullspace property holds. Then the optimal values of CCP (15) *and* ALSO-X (16) *coincide, i.e., $v^A = v^*$. Moreover, every optimal solution $(\boldsymbol{x}^*, s^*(\cdot))$ to the CCP* (15) *shares the same $s^*(\cdot)$.*

Proof.

(i) We prove the "only if" direction by contradiction. Suppose that the generalized nullspace property does not hold. Then there exists a solution $(\widehat{\boldsymbol{x}}, \widehat{s}(\cdot))$ with $\tilde{\boldsymbol{\xi}}^\top(\boldsymbol{A}\widehat{\boldsymbol{x}}) - \boldsymbol{B}\widehat{\boldsymbol{x}} - \widehat{s}(\tilde{\boldsymbol{\xi}}) = 0$ and $\widehat{s}(\tilde{\boldsymbol{\xi}}) \neq 0$ a.s. and a $\mathbb{P}-$measurable set $\mathcal{S} \subseteq \Xi$ such that $\mathbb{P}\{\tilde{\boldsymbol{\xi}}\colon \tilde{\boldsymbol{\xi}} \in \mathcal{S}\} \leq \varepsilon$ and $\mathbb{E}[|\widehat{s}(\tilde{\boldsymbol{\xi}})|\mathbb{I}(\tilde{\boldsymbol{\xi}} \in \mathcal{S})] \geq \mathbb{E}[|\widehat{s}(\tilde{\boldsymbol{\xi}})|\mathbb{I}(\tilde{\boldsymbol{\xi}} \notin \mathcal{S})]$. Since $\tilde{\boldsymbol{\xi}}^\top(\boldsymbol{A}\widehat{\boldsymbol{x}}) - \boldsymbol{B}\widehat{\boldsymbol{x}} - \widehat{s}(\tilde{\boldsymbol{\xi}}) = 0$ holds a.s., we have $\tilde{\boldsymbol{\xi}}^\top(\boldsymbol{A}\widehat{\boldsymbol{x}})\mathbb{I}(\tilde{\boldsymbol{\xi}} \in \mathcal{S}) - \boldsymbol{B}\widehat{\boldsymbol{x}} - \widehat{s}(\tilde{\boldsymbol{\xi}})\mathbb{I}(\tilde{\boldsymbol{\xi}} \in \mathcal{S}) = 0$ and $\tilde{\boldsymbol{\xi}}^\top(\boldsymbol{A}\widehat{\boldsymbol{x}})\mathbb{I}(\tilde{\boldsymbol{\xi}} \notin \mathcal{S}) - \boldsymbol{B}\widehat{\boldsymbol{x}} - \widehat{s}(\tilde{\boldsymbol{\xi}})\mathbb{I}(\tilde{\boldsymbol{\xi}} \notin \mathcal{S}) = 0$ a.s.. Then, $(\widehat{\boldsymbol{x}}, \widehat{s}(\boldsymbol{\xi})\mathbb{I}(\boldsymbol{\xi} \in \mathcal{S}))$ is not the unique optimal solution to the problem

$$\min_{\boldsymbol{x}, s(\cdot)} \left\{ \mathbb{E}\left[|s(\tilde{\boldsymbol{\xi}})|\right] : \begin{array}{l} \tilde{\boldsymbol{\xi}}^\top(\boldsymbol{A}\boldsymbol{x}) - \boldsymbol{b}^\top\boldsymbol{x} - s(\tilde{\boldsymbol{\xi}}) = \tilde{\boldsymbol{\xi}}^\top(\boldsymbol{A}\widehat{\boldsymbol{x}}) - \boldsymbol{B}\widehat{\boldsymbol{x}} - \widehat{s}(\tilde{\boldsymbol{\xi}})\mathbb{I}(\tilde{\boldsymbol{\xi}} \in \mathcal{S}), \\ \boldsymbol{c}^\top\boldsymbol{x} = \boldsymbol{c}^\top\widehat{\boldsymbol{x}}, \boldsymbol{U}^\top\boldsymbol{x} = \boldsymbol{U}^\top\widehat{\boldsymbol{x}} \end{array} \right\}.$$

Since $(\boldsymbol{0}, -\widehat{s}(\boldsymbol{\xi})\mathbb{I}(\boldsymbol{\xi} \notin \mathcal{S}))$ is a different feasible solution to the problem, whose objective value is at least as good as $(\widehat{\boldsymbol{x}}, \widehat{s}(\boldsymbol{\xi})\mathbb{I}(\boldsymbol{\xi} \in \mathcal{S}))$. This violates the presumption that all the optimal solution $(\boldsymbol{x}^*, s^*(\cdot))$ to the hinge-loss approximation (16b) has the same $s^*(\cdot)$ and satisfies $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : s^*(\tilde{\boldsymbol{\xi}})=0\} \geq 1-\varepsilon$.

To prove the "if" direction: we let $(\boldsymbol{x}, s(\cdot))$ be a feasible solution of the hinge-loss approximation (16b) that satisfies $\mathbb{P}\{\tilde{\boldsymbol{\xi}} : s(\tilde{\boldsymbol{\xi}}) \neq 0\} \leq \varepsilon$ and $(\widehat{\boldsymbol{x}}, \widehat{s}(\cdot))$ be an optimal solution of (16b) such that $s(\cdot) \neq \widehat{s}(\cdot)$. Let us denote $\mathcal{S} = \{\boldsymbol{\xi} : s(\boldsymbol{\xi}) \neq 0\}$, $Z(\tilde{\boldsymbol{\xi}}) = s(\boldsymbol{\xi}) - \widehat{s}(\boldsymbol{\xi}) \neq 0$, and $\bar{\boldsymbol{x}} = \boldsymbol{x} - \widehat{\boldsymbol{x}}$. Then we have $\tilde{\boldsymbol{\xi}}^\top(\boldsymbol{A}\bar{\boldsymbol{x}}) - \boldsymbol{B}\bar{\boldsymbol{x}} - Z(\boldsymbol{\xi}) = 0$ a.s., $\boldsymbol{c}^\top\bar{\boldsymbol{x}} = \boldsymbol{c}^\top\boldsymbol{x} - \boldsymbol{c}^\top\widehat{\boldsymbol{x}} = 0$, and $\boldsymbol{U}^\top\bar{\boldsymbol{x}} = \boldsymbol{U}^\top\boldsymbol{x} - \boldsymbol{U}^\top\widehat{\boldsymbol{x}} = \boldsymbol{0}$. Thus,

$$\mathbb{E}\left[|s(\tilde{\boldsymbol{\xi}})|\mathbb{I}(\tilde{\boldsymbol{\xi}} \in \mathcal{S})\right] - \mathbb{E}\left[|\widehat{s}(\tilde{\boldsymbol{\xi}})|\mathbb{I}(\tilde{\boldsymbol{\xi}} \in \mathcal{S})\right] \leq \mathbb{E}\left[|s(\tilde{\boldsymbol{\xi}}) - \widehat{s}(\tilde{\boldsymbol{\xi}})|\mathbb{I}(\tilde{\boldsymbol{\xi}} \in \mathcal{S})\right]$$

$$=\mathbb{E}\left[|Z(\tilde{\boldsymbol{\xi}})|\mathbb{I}(\tilde{\boldsymbol{\xi}}\in\mathcal{S})\right]<\mathbb{E}\left[|Z(\tilde{\boldsymbol{\xi}})|\mathbb{I}(\tilde{\boldsymbol{\xi}}\notin\mathcal{S})\right]=\mathbb{E}\left[|\widehat{s}(\tilde{\boldsymbol{\xi}})|\mathbb{I}(\tilde{\boldsymbol{\xi}}\notin\mathcal{S})\right],$$

the first inequality is due to the triangular inequality and the second inequality is based on the generalized nullspace property.

Therefore, we get

$$\mathbb{E}\left[|s(\tilde{\boldsymbol{\xi}})|\right]=\mathbb{E}\left[|s(\tilde{\boldsymbol{\xi}})|\mathbb{I}(\tilde{\boldsymbol{\xi}}\in\mathcal{S})\right]<\mathbb{E}\left[|\widehat{s}(\tilde{\boldsymbol{\xi}})|\right],$$

which is a contradiction to the optimality of $\widehat{s}(\cdot)$.

(ii) This follows directly from Part (i) by letting $t=v^*$. □

A.6 Proof of Theorem 3

Theorem 3 *(A generalization of proposition 12 in Ahmed et al. 2017) Suppose that $g(\boldsymbol{x},\boldsymbol{\xi})\colon \mathcal{X}\times\Xi\to\mathbb{R}_-\cup\{M\}$, where $M\in\mathbb{R}_{++}$ is a positive constant, the optimal value of* ALSO-X (7) *coincides with that of CCP* (1)*.*

Proof. Since $v^A\geq v^*$, thus it suffices to show that $v^A\leq v^*$. In fact, we claim that for any $t\geq v^*$, there exists an optimal solution of the hinge-loss approximation (7b) which satisfies the chance constraint (7c). We prove it by contradiction. Suppose the statement is not true. That is, there exists a $t\geq v^*$ such that an optimal solution $(\bar{\boldsymbol{x}},\bar{s}(\cdot))$ of the hinge-loss approximation (7b) violates the chance constraint (i.e., $\mathbb{P}\{\tilde{\boldsymbol{\xi}}\colon \bar{s}(\tilde{\boldsymbol{\xi}})>0\}>\varepsilon$). According to the definition of random function $g(\cdot,\cdot)$, we know that $\mathbb{P}\{\tilde{\boldsymbol{\xi}}:\bar{s}(\tilde{\boldsymbol{\xi}})\in\{0,M\}\}=1$. Let $\bar{s}(\boldsymbol{\xi})=M\mathbb{I}\{g(\bar{\boldsymbol{x}},\boldsymbol{\xi})>0\}$. According to the properties of the measurable functions (see, e.g., section 3.1 in Royden and Fitzpatrick 1988), $\bar{s}(\tilde{\boldsymbol{\xi}})$ is measurable. Thus, we have $\mathbb{E}[\bar{s}(\tilde{\boldsymbol{\xi}})]\geq M\mathbb{P}\{\bar{s}(\tilde{\boldsymbol{\xi}})>0\}>M\varepsilon$.

On the other hand, let $(\boldsymbol{x}^*,s^*(\cdot),z^*(\cdot))$ be an optimal solution of CCP (3) with optimal value v^*. Define $\widehat{s}(\boldsymbol{\xi})=M-Mz^*(\boldsymbol{\xi})$, and we have

$$\mathbb{E}[\widehat{s}(\tilde{\boldsymbol{\xi}})]=M-M\mathbb{E}[z^*(\tilde{\boldsymbol{\xi}})]\leq M-M(1-\varepsilon).$$

Clearly, $(\boldsymbol{x}^*,\widehat{z}(\cdot))$ is feasible to the hinge-loss approximation (7b) with an objective value $\mathbb{E}[\widehat{s}(\tilde{\boldsymbol{\xi}})]\leq M\varepsilon$, which contradicts the optimality of $(\bar{\boldsymbol{x}},\bar{s}(\cdot))$. □

A.7 Proof of Theorem 4

Theorem 4 *For the single linear CCP* (9) *under an elliptical distribution, the* ALSO-X (10) *provides an optimal solution to CCP* (9)*, provided that (i) $\mathcal{X}\subseteq\{\boldsymbol{x}:\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}=C\}$, where C is a positive constant; or (ii) $\mathcal{X}\subseteq\{\boldsymbol{x}:b_1(\boldsymbol{x})-\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x})=C\}$, where C is an arbitrary constant.*

Proof. We split the proof into two parts by checking two conditions separately.

(i) Suppose that $\mathcal{X} \subseteq \{\boldsymbol{x} : \sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x})} = C\}$, where C is a positive constant. Then for a given t, the hinge-loss approximation (10b) becomes

$$\begin{aligned}(\boldsymbol{x}^*, \alpha^*) \in \underset{\boldsymbol{x}\in\mathcal{X},\alpha}{\arg\min} \Big\{ & C\left(\overline{G}(\alpha^2/2) - \alpha + \alpha\Phi(\alpha)\right) : \boldsymbol{c}^\top \boldsymbol{x} \le t, \\ & (b_1(\boldsymbol{x}) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x}))/C = \alpha \Big\}.\end{aligned} \tag{37}$$

Let $f(\alpha) = \overline{G}(\alpha^2/2) - \alpha + \alpha\Phi(\alpha)$. Since its derivative is

$$\partial f(\alpha)/\partial\alpha = -k\alpha\widehat{g}(\alpha^2/2) - 1 + \Phi(\alpha) + k\alpha\widehat{g}(\alpha^2/2) = \Phi(\alpha) - 1 < 0,$$

function $f(\alpha)$ is monotone decreasing over $\alpha \in \mathbb{R}$. Thus, for any $t \ge v^*$, i.e., there exists an $\widehat{\boldsymbol{x}}$ which is feasible to CCP (9) such that

$$(b_1(\widehat{\boldsymbol{x}}) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\widehat{\boldsymbol{x}}))/C \ge \Phi^{-1}(1-\varepsilon), \boldsymbol{c}^\top \widehat{\boldsymbol{x}} \le t.$$

Clearly, $(\widehat{\boldsymbol{x}}, \widehat{\alpha} = \Phi^{-1}(1-\varepsilon))$ is feasible to the hinge-loss approximation (37). Due to the monotonicity of the objective function, we must have $\alpha^* \ge \widehat{\alpha} = \Phi^{-1}(1-\varepsilon)$, i.e.,

$$b_1(\boldsymbol{x}^*) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x}^*) \ge \Phi^{-1}(1-\varepsilon)C.$$

Hence, we must have $\boldsymbol{x}^*$ is also feasible to CCP (9), i.e., $v^A \le v^*$. On the other hand, we always have $v^A \ge v^*$. Thus, $v^A = v^*$.

(ii) Suppose that $\mathcal{X} \subseteq \{\boldsymbol{x} : b_1(\boldsymbol{x}) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x}) = C\}$, where C is an arbitrary constant. Let us denote $\sigma = \sqrt{\boldsymbol{a}_1^\top(\boldsymbol{x})\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}$, and we have $\alpha = C/\sigma$. Then for a given t, the hinge-loss approximation (10b) becomes

$$(\boldsymbol{x}^*, \sigma^*) \in \underset{\boldsymbol{x}\in\mathcal{X},\sigma}{\arg\min} \left\{ \sigma f(C/\sigma) \colon \sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x})} = \sigma, \boldsymbol{c}^\top \boldsymbol{x} \le t \right\}, \tag{38}$$

where function $f(\cdot)$ is defined in the proof of Part (i). Taking the derivative of the objective function with respect to σ, we have

$$\frac{\partial(\sigma f(C/\sigma))}{\partial\sigma} = f(C/\sigma) - C/\sigma \frac{\partial f(\alpha)}{\partial\alpha}\Big|_{\alpha=C/\sigma} = \overline{G}(C^2/(2\sigma^2)) > 0.$$

Thus, the objective function in the hinge-loss approximation is always monotone increasing with respect to σ. Thus, for any $t \ge v^*$, i.e., there exists a $\widehat{\boldsymbol{x}}$ which is feasible to CCP (9) such that

$$C \ge \Phi^{-1}(1-\varepsilon)\sqrt{\boldsymbol{a}_1(\widehat{\boldsymbol{x}})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\widehat{\boldsymbol{x}})}, \boldsymbol{c}^\top \widehat{\boldsymbol{x}} \le t.$$

Clearly, $(\widehat{\boldsymbol{x}}, \widehat{\sigma} = \sqrt{\boldsymbol{a}_1(\widehat{\boldsymbol{x}})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\widehat{\boldsymbol{x}})})$ is feasible to the hinge-loss approximation (37). Due to the monotonicity of the objective function, we must have $\sigma^* \le \widehat{\sigma} \le C/\Phi^{-1}(1-\varepsilon)$, i.e.,

$$b_1(\boldsymbol{x}^*) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x}^*) := C \ge \Phi^{-1}(1-\varepsilon)\sqrt{\boldsymbol{a}_1(\boldsymbol{x}^*)^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x}^*)}.$$

Hence, we must have $\boldsymbol{x}^*$ is also feasible to CCP (9), i.e., $v^A \le v^*$. On the other hand, we always have $v^A \ge v^*$. Thus, $v^A = v^*$. □

A.8 Proof of Theorem 5

Theorem 5 *For the covering CCP* (18)*, the* ALSO-X (19) *yields a* $(\lfloor N\varepsilon\rfloor+1)-$*approximation guarantee, that is,* $v^A/v^*\le\lfloor N\varepsilon\rfloor+1$.

Proof. First, without loss of generality, we assume $v^{rel}>0$; otherwise, we must have $v^{rel}=v^A=v^*=0$ due to the covering structure. Next, we split the proof into three steps.

Step 1. Let $\alpha,\{\boldsymbol{\beta}_i\}_{i\in[N]}$ be the dual variables of the constraints of the continuous relaxation (20), respectively. Then the dual of the continuous relaxation (20) is

$$v^{rel}=\max_{\alpha\ge0,\boldsymbol{\beta}_i\ge\mathbf{0},\forall i\in[N]} \quad -\lfloor N\varepsilon\rfloor\alpha+\sum_{i\in[N]}\boldsymbol{\beta}_i^\top\boldsymbol{e}, \tag{39a}$$

$$\text{s.t.}\ \sum_{i\in[N]}\boldsymbol{A}^{i\top}\boldsymbol{\beta}_i\le\boldsymbol{c}, \tag{39b}$$

$$\boldsymbol{\beta}_i^\top\boldsymbol{e}\le\alpha, \tag{39c}$$

where the strong duality holds since the continuous relaxation (20) is always feasible.

For any $t\ge0$, let $\gamma,\{\boldsymbol{\omega}_i\}_{i\in[N]}$ be the dual variables of the constraints of the hinge-loss approximation (19b), respectively. The dual of the hinge-loss approximation (19b) is

$$v^A(t)=\max_{\gamma\ge0,\boldsymbol{\omega}_i\ge\mathbf{0},\forall i\in[N]} \quad -t\gamma+\sum_{i\in[N]}\boldsymbol{\omega}_i^\top\boldsymbol{e}, \tag{40a}$$

$$\text{s.t.}\ \sum_{i\in[N]}\boldsymbol{A}^{i\top}\boldsymbol{\omega}_i\le\gamma\boldsymbol{c}, \tag{40b}$$

$$\boldsymbol{\omega}_i^\top\boldsymbol{e}\le\frac{1}{N}, \tag{40c}$$

where the strong duality also holds since the hinge-loss approximation (19b) is always feasible.

Step 2. Next, we prove the result by contradiction. Suppose that in the hinge-loss approximation (19b), for a given $t\ge v^{rel}(\lfloor N\varepsilon\rfloor+1)>0$, there exists an optimal solution $(\boldsymbol{x}^*,\boldsymbol{s}^*)$, which is infeasible to the upper-level problem of ALSO-X (19), i.e., $|\text{supp}(\boldsymbol{s}^*)|\ge\lfloor N\varepsilon\rfloor+1$. Let $(\gamma^*,\{\boldsymbol{\omega}_i^*\}_{i\in[N]})$ denote an optimal dual solution of (40). Due to the complementary slackness, we must have

$$\boldsymbol{\omega}_i^{*\top}\boldsymbol{e}=\frac{1}{N},\forall i\in[N]:s_i^*>0. \tag{41}$$

Let $(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{s}})$ be an optimal solution to the continuous relaxation (20). According to Ahmed and Xie (2018), the scaled solution $((\lfloor N\varepsilon\rfloor+1)\widehat{\boldsymbol{x}},\min\{\lceil(\lfloor N\varepsilon\rfloor+1)\widehat{\boldsymbol{s}}\rceil,\mathbf{e}\})$ is feasible to the covering CCP (18) with objective value at most $(\lfloor N\varepsilon\rfloor+1)v^{rel}$. Thus, the scaled solution $((\lfloor N\varepsilon\rfloor+1)\widehat{\boldsymbol{x}},\min\{\lceil(\lfloor N\varepsilon\rfloor+1)\widehat{\boldsymbol{s}}\rceil,\mathbf{e}\})$ is also feasible to the hinge-loss approximation (19b) since $t\ge v^{rel}(\lfloor N\varepsilon\rfloor+1)>0$. Hence, we must have

$$0<v^A(t)=\frac{1}{N}\sum_{i\in[N]}s_i^*=-t\gamma^*+\sum_{i\in[N]}\boldsymbol{\omega}_i^{*\top}\boldsymbol{e}\le\frac{\lfloor N\varepsilon\rfloor}{N}. \tag{42}$$

According to (41), we have

$$\sum_{i\in[N]} {\boldsymbol{\omega}_i^*}^\top \boldsymbol{e} \geq \frac{1}{N}|\mathrm{supp}(\boldsymbol{s}^*)| \geq \frac{\lfloor N\varepsilon\rfloor+1}{N}.$$

Together with the second inequality in (42), we must have $\gamma^*>0$.

Also, the first inequality in (42) implies that

$$\gamma^* < \frac{\sum_{i\in[N]} {\boldsymbol{\omega}_i^*}^\top \boldsymbol{e}}{t}.$$

Step 3. Now, let us define $\widehat{\boldsymbol{\beta}}_i = \boldsymbol{\omega}_i^*/\gamma^*$ and $\widehat{\alpha}=1/(N\gamma^*)$. Clearly, $(\widehat{\alpha},\{\widehat{\boldsymbol{\beta}}_i\}_{i\in[N]})$ is feasible to the dual (39) of the continuous relaxation, whose objective is equal to

$$-\lfloor N\varepsilon\rfloor\widehat{\alpha}+\sum_{i\in[N]}\widehat{\boldsymbol{\beta}}_i^\top\boldsymbol{e} = \frac{\sum_{i\in[N]}{\boldsymbol{\omega}_i^*}^\top\boldsymbol{e}-\lfloor N\varepsilon\rfloor/N}{\gamma^*} > \frac{t(\sum_{i\in[N]}{\boldsymbol{\omega}_i^*}^\top\boldsymbol{e}-\lfloor N\varepsilon\rfloor/N)}{\sum_{i\in[N]}{\boldsymbol{\omega}_i^*}^\top\boldsymbol{e}} \geq \frac{t}{\lfloor N\varepsilon\rfloor+1}\geq v^{rel},$$

where the first inequality is due to the fact that $\gamma^* < \sum_{i\in[N]}{\boldsymbol{\omega}_i^*}^\top\boldsymbol{e}/t$, the second inequality is because function $f(x)=(x-\lfloor N\varepsilon\rfloor/N)/x$ is monotone increasing with respect to x if $x\geq\lfloor N\varepsilon\rfloor/N$, and the third one is due to the assumption that $t\geq v^{rel}(\lfloor N\varepsilon\rfloor+1)>0$. This contradicts the weak duality that $-\lfloor N\varepsilon\rfloor\widehat{\alpha}+\sum_{i\in[N]}\widehat{\boldsymbol{\beta}}_i^\top\boldsymbol{e}\leq v^{rel}$. □

A.9 Proof of Proposition 5

Proposition 5 *For the covering CCP* (18)*, the* $(\lfloor N\varepsilon\rfloor+1)-$*approximation ratio of* ALSO-X (19) *is tight, i.e., it is possible that* $v^A/v^*=\lfloor N\varepsilon\rfloor+1$.

Proof. Let us consider the following example.

Example Consider a CCP with N equiprobable scenarios (i.e., $\mathbb{P}\{\tilde{\boldsymbol{\xi}}=\boldsymbol{\xi}^i\}=1/N$), risk level $\varepsilon>1/N$, set $\mathcal{X}=\mathbb{R}_+^{\lfloor N\varepsilon\rfloor+1}$, function $g(\boldsymbol{x},\boldsymbol{\xi})=1-\boldsymbol{\xi}^\top\boldsymbol{x}$, and $\boldsymbol{\xi}^i=\boldsymbol{e}_i$ for $i\in[\lfloor N\varepsilon\rfloor+1]$, $\boldsymbol{\xi}^i=\boldsymbol{e}$ for $i\in[N]\setminus[\lfloor N\varepsilon\rfloor+1]$. ◇

In this example, the covering CCP (18) reduces to

$$v^*=\min_{\boldsymbol{x}\in\mathbb{R}_+^{\lfloor N\varepsilon\rfloor+1},\boldsymbol{z}\in\{0,1\}^N}\left\{\sum_{i\in[\lfloor N\varepsilon\rfloor+1]}x_i:\ \begin{array}{c} x_i\geq z_i,\forall i\in[\lfloor N\varepsilon\rfloor+1],\\ \sum_{j\in[\lfloor N\varepsilon\rfloor+1]}x_j\geq z_i,\forall i\in[N]\setminus[\lfloor N\varepsilon\rfloor+1],\\ \sum_{i\in[N]}z_i\geq N-\lfloor N\varepsilon\rfloor\end{array}\right\}$$

with the optimal value $v^*=1$.

The corresponding ALSO-X (19) counterpart is

$$v^A=\min_t\quad t,$$

$$\text{s.t.} \quad (\boldsymbol{x}^*, \boldsymbol{s}^*) \in \mathop{\arg\min}_{\boldsymbol{x}\in\mathbb{R}_+^{\lfloor N\varepsilon\rfloor+1}, \boldsymbol{s}\in\mathbb{R}_+^N} \left\{ \frac{1}{N}\sum_{i\in[N]} s_i : \begin{array}{c} x_i \geq 1-s_i, \forall i \in [\lfloor N\varepsilon\rfloor+1], \\ \sum_{j\in[\lfloor N\varepsilon\rfloor+1]} x_j \geq 1-s_i, \forall i \in [N]\setminus[\lfloor N\varepsilon\rfloor+1], \\ \sum_{i\in[\lfloor N\varepsilon\rfloor+1]} x_i \leq t \end{array} \right\},$$
$$\sum_{i\in[N]} \mathbb{I}(s_i^*=0) \geq N - \lfloor N\varepsilon\rfloor.$$

For any $1 \leq t < \lfloor N\varepsilon\rfloor + 1$, an optimal solution $(\boldsymbol{x}^*, \boldsymbol{s}^*)$ of the hinge-loss approximation is

$$x_j^* = \frac{t}{\lfloor N\varepsilon\rfloor+1}, \forall j \in [\lfloor N\varepsilon\rfloor+1], \quad s_i^* = \begin{cases} \frac{\lfloor N\varepsilon\rfloor+1-t}{\lfloor N\varepsilon\rfloor+1} > 0, & \text{if } i \in [\lfloor N\varepsilon\rfloor+1] \\ 0, & \text{otherwise} \end{cases}, \forall i \in [N],$$

which violates the chance constraint since $\sum_{i\in[N]} \mathbb{I}(s_i^*=0) = N - \lfloor N\varepsilon\rfloor - 1 < N - \lfloor N\varepsilon\rfloor$. On the other hand, if $t \geq \lfloor N\varepsilon\rfloor + 1$, the optimal solution $(\boldsymbol{x}^*, \boldsymbol{s}^*)$ of the hinge-loss approximation is $\boldsymbol{x}^* = \mathbf{e}, \boldsymbol{s}^* = \mathbf{0}$, which satisfies the chance constraint. Thus, in this example, we have $v^A = \lfloor N\varepsilon\rfloor + 1 = (\lfloor N\varepsilon\rfloor + 1)v^*$. This completes the proof. □

Proofs in Section 4

A.10 Proof of Proposition 6

Proposition 6 *The sequence of objective values* $\{\mathbb{E}[z^k(\tilde{\boldsymbol{\xi}})s^k(\tilde{\boldsymbol{\xi}})]\}_{k\in\mathbb{Z}_+}$ *generated by the* AM *Algorithm 3 is monotonically nonincreasing, bounded from below, and hence converges.*

Proof. At iteration $k+1$ of AM Algorithm 3, the optimality condition of problem (21a) implies that

$$\mathbb{E}\left[z^{k+1}(\tilde{\boldsymbol{\xi}})s^k(\tilde{\boldsymbol{\xi}})\right] \leq \mathbb{E}\left[z^k(\tilde{\boldsymbol{\xi}})s^k(\tilde{\boldsymbol{\xi}})\right].$$

Similarly, the optimality condition of problem (21b) implies that

$$\mathbb{E}\left[z^{k+1}(\tilde{\boldsymbol{\xi}})s^{k+1}(\tilde{\boldsymbol{\xi}})\right] \leq \mathbb{E}\left[z^{k+1}(\tilde{\boldsymbol{\xi}})s^k(\tilde{\boldsymbol{\xi}})\right].$$

Therefore, we have

$$\mathbb{E}\left[z^{k+1}(\tilde{\boldsymbol{\xi}})s^{k+1}(\tilde{\boldsymbol{\xi}})\right] \leq \mathbb{E}\left[z^{k+1}(\tilde{\boldsymbol{\xi}})s^k(\tilde{\boldsymbol{\xi}})\right] \leq \mathbb{E}\left[z^k(\tilde{\boldsymbol{\xi}})s^k(\tilde{\boldsymbol{\xi}})\right].$$

Thus, the sequence of $\mathbb{E}[z^k(\tilde{\boldsymbol{\xi}})s^k(\tilde{\boldsymbol{\xi}})]$ is monotonically nonincreasing. The fact that both $s^k(\boldsymbol{\xi})$ and $z^k(\boldsymbol{\xi})$ are nonnegative implies that $\mathbb{E}[z^k(\tilde{\boldsymbol{\xi}})s^k(\tilde{\boldsymbol{\xi}})] \geq 0$ at any iteration k. Hence, Monotone Convergence Theorem (see, e.g., thereon 7.16 in Rudin et al. 1964) implies that the sequence of objective values $\{\mathbb{E}[z^k(\tilde{\boldsymbol{\xi}})s^k(\tilde{\boldsymbol{\xi}})]\}_{k\in\mathbb{Z}_+}$ is convergent. □

Proofs in Section 5

A.11 Proof of Proposition 8

Proposition 8 *Under $\infty-$Wasserstein ambiguity set, DRCCP* (27) *is equivalent to*

$$v_\infty^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \mathbb{P}_{\tilde{\boldsymbol{\zeta}}} \left\{ \tilde{\boldsymbol{\zeta}} \colon \bar{g}_i(\boldsymbol{x}, \tilde{\boldsymbol{\zeta}}) \le 0, \forall i \in [I] \right\} \ge 1-\varepsilon \right\}, \tag{28}$$

where the convex and lower semi-continuous function $\bar{g}_i : \mathbb{R}^n \times \Xi \to \mathbb{R}$ is defined as $\bar{g}_i(\boldsymbol{x}, \boldsymbol{\zeta}) := \max_{\boldsymbol{\xi}}\{g_i(\boldsymbol{x}, \boldsymbol{\xi}) \colon \|\boldsymbol{\xi} - \boldsymbol{\zeta}\| \le \theta\}$ for each $i \in [I]$.

Proof. We first prove the following claim.

Claim 1 *For any $\mathbb{P}-$measurable function $f(\boldsymbol{\xi}) : \Xi \to \mathbb{R}$ with $\mathbb{P} \in \mathcal{P}_\infty^W$, we must have*

$$\sup_{\mathbb{P}\in\mathcal{P}_\infty^W} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{\xi}})] = \mathbb{E}_{\mathbb{P}_{\tilde{\zeta}}} \left[\sup_{\boldsymbol{\xi}} \left\{ f(\boldsymbol{\xi}) : \|\boldsymbol{\xi} - \tilde{\boldsymbol{\zeta}}\| \le \theta \right\} \right].$$

Proof. It is sufficient to prove that

$$\lim_{q\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_q^W} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{\xi}})] = \mathbb{E}_{\mathbb{P}_{\tilde{\zeta}}} \left[\sup_{\boldsymbol{\xi}} \left\{ f(\boldsymbol{\xi}) : \|\boldsymbol{\xi} - \tilde{\boldsymbol{\zeta}}\| \le \theta \right\} \right].$$

For any $q \ge 1$, according to theorem 1 in Gao and Kleywegt (2016) or theorem 1 in Blanchet and Murthy (2019), $\sup_{\mathbb{P}\in\mathcal{P}_q^W} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{\xi}})]$ can be reformulated as

$$\sup_{\mathbb{P}\in\mathcal{P}_q^W} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{\xi}})] = \min_{\lambda\ge 0} \lambda\theta^q + \mathbb{E}_{\mathbb{P}_{\tilde{\zeta}}} \left[\sup_{\boldsymbol{\xi}} \left\{ f(\boldsymbol{\xi}) - \lambda \left\| \boldsymbol{\xi} - \tilde{\boldsymbol{\zeta}} \right\|^q \right\} \right].$$

First of all, interchanging the expectation and inner supremum operator with the outer minimum operator, we arrive at the lower bound as

$$\sup_{\mathbb{P}\in\mathcal{P}_q^W} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{\xi}})] \ge \mathbb{E}_{\mathbb{P}_{\tilde{\zeta}}} \left[\sup_{\boldsymbol{\xi}} \left\{ f(\boldsymbol{\xi}) + \min_{\lambda\ge 0} \left\{ \lambda\theta^q - \lambda \left\| \boldsymbol{\xi} - \tilde{\boldsymbol{\zeta}} \right\|^q \right\} \right\} \right] := \mathbb{E}_{\mathbb{P}_{\tilde{\zeta}}} \left[\sup_{\boldsymbol{\xi}} \left\{ f(\boldsymbol{\xi}) : \|\boldsymbol{\xi} - \tilde{\boldsymbol{\zeta}}\| \le \theta \right\} \right].$$

Since the $q-$Wasserstein distance is monotone nondecreasing as q increases (according to Jensen's inequality), thus $\{\sup_{\mathbb{P}\in\mathcal{P}_q^W} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{\xi}})]\}_{q\ge 1}$ is a monotone nonincreasing sequence and is bounded from below. Thus, its limit exists and thus, we have

$$\sup_{\mathbb{P}\in\mathcal{P}_\infty^W} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{\xi}})] := \lim_{q\to\infty} \sup_{\mathbb{P}\in\mathcal{P}_q^W} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{\xi}})] \ge \mathbb{E}_{\mathbb{P}_{\tilde{\zeta}}} \left[\max_{\boldsymbol{\xi}} \left\{ f(\boldsymbol{\xi}) : \|\boldsymbol{\xi} - \tilde{\boldsymbol{\zeta}}\| \le \theta \right\} \right].$$

On the other hand, let us define a random vector $\tilde{\boldsymbol{\xi}}$ as $\tilde{\boldsymbol{\xi}} \in \arg\max_{\boldsymbol{\xi}}\{f(\boldsymbol{\xi}) : \|\boldsymbol{\xi} - \tilde{\boldsymbol{\zeta}}\| \le \theta\}$. According to the definition, we have the $\infty-$Wasserstein distance $W_\infty(\mathbb{P}_{\tilde{\boldsymbol{\xi}}}, \mathbb{P}_{\tilde{\zeta}})$ no larger than θ. That is, $\mathbb{P}_{\tilde{\boldsymbol{\xi}}} \in \mathcal{P}_\infty^W$ and

$$\mathbb{E}_{\mathbb{P}_{\tilde{\boldsymbol{\xi}}}}[f(\tilde{\boldsymbol{\xi}})] = \mathbb{E}_{\mathbb{P}_{\tilde{\zeta}}} \left[\max_{\boldsymbol{\xi}} \left\{ f(\boldsymbol{\xi}) : \|\boldsymbol{\xi} - \tilde{\boldsymbol{\zeta}}\| \le \theta \right\} \right].$$

Thus, the lower bound $\mathbb{E}_{\mathbb{P}_{\tilde{\zeta}}}[\max_{\boldsymbol{\xi}}\{f(\boldsymbol{\xi}) : \|\boldsymbol{\xi} - \tilde{\boldsymbol{\zeta}}\| \le \theta\}]$ of $\sup_{\mathbb{P}\in\mathcal{P}_\infty^W} \mathbb{E}_{\mathbb{P}}[f(\tilde{\boldsymbol{\xi}})]$ is attainable. This concludes the proof. ◇

For the distributionally robust chance constraint in (27), according to Claim 1, we have

$$\inf_{\mathbb{P}\in\mathcal{P}_\infty^W} \mathbb{P}\left\{\tilde{\boldsymbol{\xi}}\colon g_i(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\le 0,\forall i\in[I]\right\} = \mathbb{P}_{\tilde{\boldsymbol{\zeta}}}\left\{\tilde{\boldsymbol{\zeta}}\colon g_i(\boldsymbol{x},\tilde{\boldsymbol{\xi}})\le 0,\forall\|\tilde{\boldsymbol{\xi}}-\tilde{\boldsymbol{\zeta}}\|\le\theta,\forall i\in[I]\right\}.$$

Recall the definition of $\bar{g}_i(\boldsymbol{x},\boldsymbol{\zeta})$, we have the following equivalent representation of (27)

$$\mathbb{P}_{\tilde{\boldsymbol{\zeta}}}\left\{\tilde{\boldsymbol{\zeta}}\colon \bar{g}_i(\boldsymbol{x},\tilde{\boldsymbol{\zeta}})\le 0,\forall i\in[I]\right\}\ge 1-\varepsilon.$$

This completes the proof. □

A.12 Proof of Proposition 9

Proposition 9 *Under $\infty-$Wasserstein ambiguity set, we have*

(i) the worst-case ALSO-X (29) *is equivalent to*

$$\begin{aligned} v_\infty^A = \min_t \quad & t, \\ \text{s.t.} \quad & \boldsymbol{x}^*\in\arg\min_{\boldsymbol{x}\in\mathcal{X}}\left\{\mathbb{E}_{\mathbb{P}}\left[\max_{i\in[I]}\bar{g}_i(\boldsymbol{x},\tilde{\boldsymbol{\xi}})_+\right]\colon \boldsymbol{c}^\top\boldsymbol{x}\le t\right\}, \\ & \mathbb{P}_{\tilde{\boldsymbol{\zeta}}}\left\{\tilde{\boldsymbol{\zeta}}\colon \bar{g}_i(\boldsymbol{x}^*,\tilde{\boldsymbol{\zeta}})\le 0,\forall i\in[I]\right\}\ge 1-\varepsilon; \end{aligned} \tag{31}$$

(ii) the worst-case CVaR *approximation* (30) *is equivalent to*

$$v_\infty^{\mathrm{CVaR}} = \min_{\boldsymbol{x}\in\mathcal{X}}\left\{\boldsymbol{c}^\top\boldsymbol{x}\colon \min_\beta\left[\beta+\frac{1}{\varepsilon}\mathbb{E}_{\mathbb{P}_{\tilde{\boldsymbol{\zeta}}}}\left\{\max_{i\in[I]}\left\{\bar{g}_i(\boldsymbol{x},\tilde{\boldsymbol{\zeta}})\right\}-\beta\right\}_+\right]\le 0\right\}. \tag{32}$$

Proof. We split the proof into two parts by proving the reformulations of the worst-case ALSO-X and the worst-case CVaR approximation, separately.

(i) According to Claim 1 in Appendix A.11, we can rewrite the objective function of the worst-case hinge-loss approximation (29b) under $\infty-$Wasserstein ambiguity set as

$$\sup_{\mathbb{P}\in\mathcal{P}_\infty^W}\mathbb{E}_{\mathbb{P}}\left[\max_{i\in[I]}g_i(\boldsymbol{x},\boldsymbol{\xi})_+\right] = \mathbb{E}_{\mathbb{P}_{\tilde{\boldsymbol{\zeta}}}}\left\{\max_{\boldsymbol{\xi}}\max_{i\in[I]}\{g_i(\boldsymbol{x},\boldsymbol{\xi})\}_+\colon \left\|\boldsymbol{\xi}-\tilde{\boldsymbol{\zeta}}\right\|\le\theta\right\}.$$

Interchanging the maximum operators, we have

$$\sup_{\mathbb{P}\in\mathcal{P}_\infty^W}\mathbb{E}_{\mathbb{P}}\left[\max_{i\in[I]}g_i(\boldsymbol{x},\boldsymbol{\xi})_+\right] = \mathbb{E}_{\mathbb{P}_{\tilde{\boldsymbol{\zeta}}}}\left\{\max_{i\in[I]}\max_{\boldsymbol{\xi}}\{g_i(\boldsymbol{x},\boldsymbol{\xi})\}_+\colon \left\|\boldsymbol{\xi}-\tilde{\boldsymbol{\zeta}}\right\|\le\theta\right\}.$$

According to the definition of functions $\{\bar{g}_i(\cdot,\cdot)\}_{i\in[I]}$, we can further rewrite the objective function as

$$\sup_{\mathbb{P}\in\mathcal{P}_\infty^W}\mathbb{E}_{\mathbb{P}}\left[\max_{i\in[I]}g_i(\boldsymbol{x},\boldsymbol{\xi})_+\right] = \mathbb{E}_{\mathbb{P}_{\tilde{\boldsymbol{\zeta}}}}\left[\max_{i\in[I]}\bar{g}_i(\boldsymbol{x},\boldsymbol{\zeta})_+\right].$$

Thus, the worst-case hinge-loss approximation (29b) is equivalent to

$$\boldsymbol{x}^*\in\arg\min_{\boldsymbol{x}\in\mathcal{X}}\left\{\mathbb{E}_{\mathbb{P}}\left[\max_{i\in[I]}\bar{g}_i(\boldsymbol{x},\tilde{\boldsymbol{\xi}})_+\right]\colon \boldsymbol{c}^\top\boldsymbol{x}\le t\right\}.$$

According to Proposition 8, the worst-case chance constraint is equivalent to the regular chance constraint (28). Therefore, we arrive at the reformulation (31).

(ii) In the worst-case CVaR approximation (30), we can interchange the supremum operator with the minimum one, since $\mathbb{P}_{\tilde{\boldsymbol{\zeta}}}$ is sub-Gaussian, the definition of $\infty-$Wasserstein ambiguity set shows that for any $t \geq 0$ and $\mathbb{P} \in \mathcal{P}_\infty^W$, we have

$$\mathbb{P}\{\tilde{\boldsymbol{\xi}} : \|\tilde{\boldsymbol{\xi}}\| \geq t+\theta\} = \mathbb{P}_{\tilde{\zeta}}\{\tilde{\boldsymbol{\zeta}} : \|\tilde{\boldsymbol{\xi}}\| \geq t+\theta, \|\tilde{\boldsymbol{\xi}} - \tilde{\boldsymbol{\zeta}}\| \leq \theta\} \leq \mathbb{P}_{\tilde{\zeta}}\{\tilde{\boldsymbol{\zeta}} : \|\tilde{\boldsymbol{\zeta}}\| \geq t\} \leq C_1 e^{-C_2 t^2},$$

for some positive constants $C_1, C_2 > 0$, and thus $\mathcal{P}_\infty^W$ is weakly compact. Thus, according to corollary Terkelsen (1972), the worst-case CVaR approximation is equivalent to

$$v_\infty^{\mathrm{CVaR}} = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \min_{\beta} \left[\beta + \frac{1}{\varepsilon} \sup_{\mathbb{P}\in\mathcal{P}_\infty^W} \mathbb{E}_{\mathbb{P}} \left\{ \max_{i\in[I]} \{g_i(\boldsymbol{x}, \boldsymbol{\xi})\} - \beta \right\}_+ \right] \leq 0 \right\}.$$

According to Claim 1 in Appendix A.11, the worst-case CVaR approximation becomes

$$v_\infty^{\mathrm{CVaR}} = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \min_{\beta} \left[\beta + \frac{1}{\varepsilon} \mathbb{E}_{\mathbb{P}_{\tilde{\zeta}}} \left[\max_{\boldsymbol{\xi}} \left\{ \max_{i\in[I]} \{g_i(\boldsymbol{x}, \boldsymbol{\xi})\} - \beta \right\}_+ : \left\| \boldsymbol{\xi} - \tilde{\boldsymbol{\zeta}} \right\| \leq \theta \right] \right] \leq 0 \right\}.$$

Interchanging the maximum operators and taking the optimization over $\boldsymbol{\xi}$, we have

$$v_\infty^{\mathrm{CVaR}} = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \min_{\beta} \left[\beta + \frac{1}{\varepsilon} \mathbb{E}_{\mathbb{P}_{\tilde{\zeta}}} \left\{ \max_{i\in[I]} \{\bar{g}_i(\boldsymbol{x}, \boldsymbol{\zeta})\} - \beta \right\}_+ \right] \leq 0 \right\}.$$

This completes the proof. □

A.13 Proof of Theorem 6

Theorem 6 *For DRCCP with $\infty-$Wasserstein ambiguity set, the worst-case* ALSO-X *outperforms the worst-case* CVaR *approximation.*

Proof. According to Proposition 9, the worst-case ALSO-X and the worst-case CVaR approximation correspond to the same regular chance constrained program. Based on Theorem 1, we know that for a regular CCP, ALSO-X yields a better solution than that of CVaR approximation. Thus, the worst-case ALSO-X can return a better solution than that of the worst-case CVaR approximation for DRCCP under $\infty-$Wasserstein ambiguity set. □

A.14 Proof of Proposition 10

Proposition 10 *Suppose that the reference distribution $\mathbb{P}_{\tilde{\zeta}}$ is elliptical $\mathbb{P}_{\mathrm{E}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{g})$, and the norm defining the Wasserstein distance is the Mahalanobis norm associated with the positive definite matrix $\boldsymbol{\Sigma}$, i.e., $\|\boldsymbol{y}\| = \sqrt{\boldsymbol{y}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}}$, for some $\boldsymbol{y} \in \mathbb{R}^n$. If $I = 1$ and the random function $g_1(\boldsymbol{x}, \boldsymbol{\xi}) = \boldsymbol{\xi}^\top \boldsymbol{a}_1(\boldsymbol{x}) - b_1(\boldsymbol{x})$, then the worst-case* ALSO-X (29) *provides an optimal solution to DRCCP* (27) *under $\infty-$Wasserstein ambiguity set if (i) $\mathcal{X} \subseteq \{\boldsymbol{x} : \sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x})} = C\}$, where C is a positive constant; or (ii) $\mathcal{X} \subseteq \{\boldsymbol{x} : b_1(\boldsymbol{x}) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x}) = C\}$, where C is an arbitrary constant.*

Proof. According to Proposition 4 and Proposition 9, for a given t, the worst-case hinge-loss approximation under $\infty-$Wasserstein ambiguity set is equivalent to

$$(\boldsymbol{x}^*,\alpha^*)\in\mathop{\arg\min}_{\boldsymbol{x}\in\mathcal{X},\alpha}\left\{\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}\,[f(\alpha-\theta)]:\boldsymbol{c}^\top\boldsymbol{x}\le t,\frac{b_1(\boldsymbol{x})-\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x})}{\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}}=\alpha\right\},\tag{43a}$$

where $f(\alpha)=\overline{G}(\alpha^2/2)-\alpha+\alpha\Phi(\alpha)$.

Next, we split the proof into two parts by checking two sufficient conditions separately.

(i) Suppose that $\mathcal{X}\subseteq\{\boldsymbol{x}:\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}=C\}$, where C is a positive constant. Then, the worst-case hinge-loss approximation can be simplified as

$$(\boldsymbol{x}^*,\alpha^*)\in\mathop{\arg\min}_{\boldsymbol{x}\in\mathcal{X},\alpha}\left\{F_1(\alpha):=C\,[f(\alpha-\theta)]:\boldsymbol{c}^\top\boldsymbol{x}\le t,\frac{b_1(\boldsymbol{x})-\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x})}{C}=\alpha\right\}.\tag{43b}$$

Since the first-order derivative of $F_1(\cdot)$ is

$$\frac{\partial F_1(\alpha)}{\partial\alpha}=C(\Phi(\alpha-\theta)-1)<0,$$

function $F_1(\alpha)$ is monotone decreasing over $\alpha\in\mathbb{R}$.

According to problem (9) and Proposition 8, we can rewrite DRCCP (27) as

$$v_\infty^*=\min_{\boldsymbol{x}\in\mathcal{X}}\left\{\boldsymbol{c}^\top\boldsymbol{x}\colon b_1(\boldsymbol{x})-\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x})\ge(\Phi^{-1}(1-\varepsilon)+\theta)\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}\right\}.\tag{43c}$$

Thus, for any $t\ge v_\infty^*$, there exists a feasible solution $\bar{\boldsymbol{x}}$ to DRCCP (43c), such that

$$\boldsymbol{\mu}^\top\boldsymbol{a}_1(\bar{\boldsymbol{x}})-b_1(\bar{\boldsymbol{x}})+(\Phi^{-1}(1-\varepsilon)+\theta)C\le0,\boldsymbol{c}^\top\bar{\boldsymbol{x}}\le t.$$

Let $\alpha'=1/C(b_1(\bar{\boldsymbol{x}})-\boldsymbol{\mu}^\top\boldsymbol{a}_1(\bar{\boldsymbol{x}}))\ge\Phi^{-1}(1-\varepsilon)+\theta$. Then, $(\bar{\boldsymbol{x}},\alpha')$ is feasible to the worst-case hinge-loss approximation (43b).

Due to the monotonicity of the objective function $F_1(\cdot)$, we must have $\alpha^*\ge\alpha'\ge\Phi^{-1}(1-\varepsilon)+\theta$, i.e.,

$$\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x}^*)-b_1(\boldsymbol{x}^*)+C\alpha^*\le0.$$

Hence, we must have $\boldsymbol{x}^*$ is also feasible to DRCCP (43c). This implies that the optimal value of the worst-case ALSO-X (29) must be $v_\infty^A\le v_\infty^*$. On the hand, we always have $v_\infty^A\ge v^*$. Thus, $v_\infty^A=v_\infty^*$.

(ii) Suppose that $\mathcal{X}\subseteq\{\boldsymbol{x}:b_1(\boldsymbol{x})-\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x})=C\}$, where C is an arbitrary constant. Let us denote $\sigma=\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}$. Then the worst-case hinge-loss approximation can be simplified as

$$(\boldsymbol{x}^*,\widehat{\alpha}^*,\sigma^*)\in\mathop{\arg\min}_{\boldsymbol{x}\in\mathcal{X},\widehat{\alpha},\sigma}\left\{F_2(\sigma,\widehat{\alpha}):=\sigma f(\widehat{\alpha})\colon\boldsymbol{c}^\top\boldsymbol{x}\le t,\widehat{\alpha}=\frac{C}{\sigma},\sigma=\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}\right\}.\tag{43d}$$

The first-order derivative of $F_2(\cdot,\cdot)$ with respect to σ is

$$\frac{\partial F_2(\sigma,\widehat{\alpha})}{\partial \sigma} = f(\widehat{\alpha}) + \sigma \frac{f(\widehat{\alpha})}{\partial \widehat{\alpha}} \frac{\partial \widehat{\alpha}}{\partial \sigma} = f(\widehat{\alpha}) + (1-\Phi(\widehat{\alpha}))\frac{C}{\sigma} > 0.$$

Thus, function $F_2(\sigma,\widehat{\alpha})$ is monotone increasing over $\sigma \in \mathbb{R}_+$. Thus, for any $t \geq v_\infty^*$, i.e., there exists a feasible solution $\widehat{\boldsymbol{x}}$ to DRCCP (28) such that

$$C \geq (\Phi^{-1}(1-\varepsilon)+\theta)\sqrt{\boldsymbol{a}_1(\widehat{\boldsymbol{x}})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\widehat{\boldsymbol{x}})}, \boldsymbol{c}^\top \widehat{\boldsymbol{x}} \leq t.$$

Let $\widehat{\sigma} = \sqrt{\boldsymbol{a}_1(\widehat{\boldsymbol{x}})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\widehat{\boldsymbol{x}})}$ and $\widehat{\alpha}' = (C/\widehat{\sigma})$. Then, $(\widehat{\boldsymbol{x}}, \widehat{\alpha}', \widehat{\sigma})$ is feasible to the worst-case hinge-loss approximation (43d). Due to the monotonicity of the objective function $F_2(\cdot,\cdot)$ with respect to σ, we must have $\sigma^* \leq \widehat{\sigma} \leq C/(\Phi^{-1}(1-\varepsilon)+\theta)$, i.e.,

$$b_1(\boldsymbol{x}^*) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x}^*) := C \geq (\Phi^{-1}(1-\varepsilon)+\theta)\sigma^* = (\Phi^{-1}(1-\varepsilon)+\theta)\sqrt{\boldsymbol{a}_1(\boldsymbol{x}^*)^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x}^*)}.$$

Hence, we must have $\boldsymbol{x}^*$ is also feasible to DRCCP (28), i.e., $v_\infty^A \leq v_\infty^*$. On the other hand, we always have $v_\infty^A \geq v_\infty^*$. Thus, $v_\infty^A = v_\infty^*$. □

Appendix B. Examples

Example 7 Consider a CCP with 3 equiprobable scenarios (i.e., $N=3$ and $\mathbb{P}\{\tilde{\xi}=\xi^i\}=1/N$), risk level $\varepsilon=1/2$, set $\mathcal{X}=\mathbb{R}_+$, function $g(\boldsymbol{x},\xi)=-x+\xi$, and $\xi^1=3$, $\xi^2=2$, $\xi^3=1$. Then the optimal solution of this CCP (1) can be obtained by solving the following mixed-integer linear program

$$v^*=\min_{x\geq 0,\boldsymbol{z}}\left\{x\colon x\geq 3z_1, x\geq 2z_2, x\geq z_3, \sum_{i\in[3]} z_i\geq 2, \boldsymbol{z}\in\{0,1\}^3\right\}.$$

Its ALSO-X counterpart admits the following form

$$v^A=\min_t\Bigg\{t\colon \sum_{i\in[3]}\mathbb{I}(s_i^*=0)\geq 2,\\ (x^*,\boldsymbol{s}^*)\in\operatorname*{arg\,min}_{x\geq 0,\boldsymbol{s}\geq\boldsymbol{0}}\left\{\frac{1}{3}\sum_{i\in[3]}s_i\colon x\geq 3-s_1, x\geq 2-s_2, x\geq 1-s_3, x\leq t\right\}\Bigg\}.$$

The CVaR approximation is

$$v^{\mathrm{CVaR}}=\min_{x\geq 0,\beta\leq 0,\boldsymbol{s}}\left\{x\colon 3-x\leq s_1, 2-x\leq s_2, 1-x\leq s_3, (s_1+s_2+s_3)/3-\beta/2\leq 0, s_i\geq\beta, \forall i\in[3]\right\}.$$

By the straightforward calculation, we obtain $v^*=2$, $v^A=2$, and $v^{\mathrm{CVaR}}=8/3$. Figure 2 illustrates their relationships, where the dotted line segment on the x-axis represents the feasible region. When $t=8/3$, the optimal solution from the hinge-loss approximation (7b) is $s_1^*=1/3, s_2^*=s_3^*=0$, which means the second constraint and the third constraint are satisfied, while the first constraint is violated. The support size of $\boldsymbol{s}^*$ is 1, i.e., $|\mathrm{supp}(\boldsymbol{s}^*)|=1$, so the current solution is feasible to the upper-level problem in ALSO-X (7) and we decrease t. Finally, we can show that ALSO-X has an optimal $t^*=2$. ◇

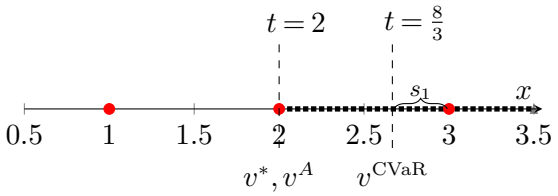

Figure 2 Illustration of Example 7

.

Example 8 Consider a CCP with 3 equiprobable scenarios (i.e., $N=3$, $\mathbb{P}\{\tilde{\boldsymbol{\xi}}=\boldsymbol{\xi}^i\}=1/N$), risk level $\varepsilon=1/3$, set $\mathcal{X}=\{0,1\}^2$, function $g(\boldsymbol{x},\boldsymbol{\xi})=1-\boldsymbol{\xi}^\top\boldsymbol{x}$, and $\boldsymbol{\xi}^1=(1,0)^\top$, $\boldsymbol{\xi}^2=(0,1)^\top$, $\boldsymbol{\xi}^3=(1,1)^\top$.

The optimal solution of this CCP can be obtained by solving the following mixed-integer linear program

$$v^* = \min_{\boldsymbol{x}\in\{0,1\}^2, \boldsymbol{z}\in\{0,1\}^3} \left\{ x_1 + 2x_2 \colon x_1 \geq z_1, x_2 \geq z_2, x_1 + x_2 \geq z_3, \sum_{i\in[3]} z_i \geq 2 \right\}$$

with the optimal value $v^* = 1$.

The corresponding ALSO-X (7) is

$$v^A = \min_t \Bigg\{ t \colon \sum_{i\in[3]} \mathbb{I}(s_i^* = 0) \geq 2, \\ (\boldsymbol{x}^*, \boldsymbol{s}^*) \in \operatorname*{arg\,min}_{\boldsymbol{x}\in\{0,1\}^2, \boldsymbol{s}\geq \boldsymbol{0}} \left\{ \frac{1}{3} \sum_{i\in[3]} s_i \colon x_1 \geq 1 - s_1, x_2 \geq 1 - s_2, x_1 + x_2 \geq 1 - s_3, x_1 + 2x_2 \leq t \right\} \Bigg\}.$$

For any $1 < t < 3$, the optimal solution is $x_1^* = 1$, $x_2^* = 0$, $s_1^* = s_3^* = 0$, $s_2^* = 1 > 0$, the support size of $\boldsymbol{s}^*$ is 1. Thus, this solution is feasible to CCP and we can decrease t until $t = 1$. Thus, the optimal value of ALSO-X is also $v^A = 1 = v^*$. ◇

Example 9 Consider a single linear CCP with a Gaussian distribution (i.e., $\tilde{\boldsymbol{\xi}} \sim \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$) with $n = 2, \bar{\boldsymbol{\mu}} = [2, 1]^\top$, $\bar{\boldsymbol{\Sigma}} = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, risk level $\varepsilon = 0.05$, set $\mathcal{X} = \mathbb{R}^2$ and function $g(\boldsymbol{x}, \boldsymbol{\xi}) = -1 + \boldsymbol{\xi}^\top \boldsymbol{x}$. This example violates both conditions in Theorem 4. We show that $v^A > v^*$.

First, in this example, CCP (9) becomes

$$v^* = \min_{\boldsymbol{x}\in\mathbb{R}^2} \left\{ -x_1 - 3x_2 : 1 - 2x_1 - x_2 \geq \Phi^{-1}(1-\varepsilon)\sqrt{x_1^2 + x_2^2} \right\},$$

and its approximate optimal value is $v^* = -1.55432$ with error bound $[-10^{-7}, 10^{-7}]$.

The corresponding ALSO-X (10) is

$$v^A = \min_t \Bigg\{ t \colon 1 - 2x_1^* - x_2^* \geq \Phi^{-1}(1-\varepsilon)\sqrt{(x_1^*)^2 + (x_2^*)^2} \\ (x_1^*, x_2^*) \in \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^2} \Bigg\{ \sqrt{x_1^2 + x_2^2} \Bigg[\varphi\Big(\frac{1 - 2x_1 - x_2}{\sqrt{x_1^2 + x_2^2}}\Big) - \frac{1 - 2x_1 - x_2}{\sqrt{x_1^2 + x_2^2}} \\ + \frac{1 - 2x_1 - x_2}{\sqrt{x_1^2 + x_2^2}} \Phi\Big(\frac{1 - 2x_1 - x_2}{\sqrt{x_1^2 + x_2^2}}\Big) \Bigg] : -x_1 - 3x_2 \leq t \Bigg\} \Bigg\}.$$

Suppose that $t = -1.42$, an approximate optimal solution of the hinge-loss approximation (10b) is $x_1^* \approx -0.375511$ with error bound $[-10^{-6}, 10^{-7}]$ and $x_2^* \approx 0.598504$ with error bound $[-3 \times 10^{-7}, 10^{-7}]$. We see that any possible solution within the error box, i.e.,

$$\max_{x_1, x_2} \left\{ 1 - 2x_1 - x_2 : x_1 - x_1^* \in [-10^{-6}, 10^{-7}], x_2 - x_2^* \in [-3 \times 10^{-7}, 10^{-7}] \right\} < 1.1526 \\ < 1.1620 < \max_{x_1, x_2} \left\{ \Phi^{-1}(1-\varepsilon)\sqrt{(x_1)^2 + (x_2)^2} : x_1 - x_1^* \in [-10^{-6}, 10^{-7}], x_2 - x_2^* \in [-3 \times 10^{-7}, 10^{-7}] \right\}.$$

Therefore, we must have $v^A \geq -1.42 > -1.55 \geq v^*$, i.e., the solution from ALSO-X (10) is not exactly optimal to the CCP. ◇

Example 10 Consider a CCP with 3 equiprobable scenarios (i.e., $\mathbb{P}\{\tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}^i\} = 1/3$ for each $i \in [3]$), risk level $\varepsilon = 1/3$, set $\mathcal{X} = \mathbb{R}^2_+$, function $g(\boldsymbol{x}, \boldsymbol{\xi}) = 1 - \boldsymbol{\xi}^\top \boldsymbol{x}$, and $\boldsymbol{\xi}^1 = (1,0)^\top$, and $\boldsymbol{\xi}^2 = \boldsymbol{\xi}^3 = (1,1)^\top$. In this case, the CCP is equivalent to the following mixed-integer linear program

$$v^* = \min_{\boldsymbol{x}\in\mathbb{R}^2_+, \boldsymbol{z}\in\{0,1\}^3} \{3x_1 + 2x_2 \colon x_1 \geq z_1, x_1 + x_2 \geq z_2, x_1 + x_2 \geq z_3, z_1 + z_2 + z_3 \geq 2\}$$

with optimal value $v^* = 2$.

The corresponding ALSO-X (7) is

$$\begin{aligned} v^A = \min_t \Bigg\{ t \colon & \sum_{i\in[3]} \mathbb{I}(s_i^* = 0) \geq 2, \\ (\boldsymbol{x}^*, \boldsymbol{s}^*) \in & \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^2_+, \boldsymbol{s}\in\mathbb{R}^3_+} \left\{ \frac{1}{3}\sum_{i\in[3]} s_i \colon \begin{array}{l} x_1 \geq 1 - s_1, x_1 + x_2 \geq 1 - s_2, \\ x_1 + x_2 \geq 1 - s_3, 3x_1 + 2x_2 \leq t \end{array} \right\} \Bigg\}, \end{aligned}$$

with the optimal $v^A = 3$.

For any $t \in [2, 3)$, an optimal solution to the hinge-loss approximation is $x_1^* = t/3, x_2^* = 0, s_1^* = s_2^* = s_3^* = 1 - t/3$. Invoking the AM Algorithm 3 with initial $s_i^0 = s_i^*$ for each $i \in [3]$, we see that $(\boldsymbol{x}^*, \boldsymbol{s}^*, \boldsymbol{z}^*)$ with $z_1^* = 1, z_2^* = 1, z_3^* = 0$ is a stationary point of the AM Algorithm 3. Therefore, in this example, ALSO-X+ Algorithm 4 with the tolerance $\delta_1 = 0$ provides the same solution as ALSO-X (7), and both fail to find an optimal solution of the CCP. ◇

Appendix C. An Illustration of ALSO-X, CVaR Approximation, and ALSO-X+ Algorithm

We use Example 3 to provide a simple illustration of ALSO-X, CVaR approximation, and ALSO-X+ algorithm. The results are shown in Figure 3, where the non-convex shaded region denotes the feasible region of the corresponding CCP studied in Example 3, and points D and E are its optimal solutions with the optimal value $v^* = 0.5$. Point F is the best solution from CVaR approximation, which is quite far away from the true optimal solution.

Given $t = 0.6$, the hinge-loss approximation (7b) is to minimize the average of violations for all constraints. Due to the symmetry of the random parameters, we see that the interval between point A and point C is the set of its optimal solutions. If one were unlucky and picked any solution inside the interval (e.g., point B) rather than the boundary points, such a choice would end up with an infeasible solution to the CCP. On the contrary, ALSO-X+ Algorithm 4 with the tolerance $\delta_1 = 0$ breaks the symmetry by circumventing the infeasible solutions like point B, and provide a better solution. For example, when $t = 0.6$ and ALSO-X+ Algorithm 4 starts at point B, it selects the two smallest constraint violations and move the solution to either point A or point C, which is feasible to the CCP.

Hence, in Example 3, we show that ALSO-X+ algorithm can find an optimal solution, while both the CVaR approximation and ALSO-X (7) may not.

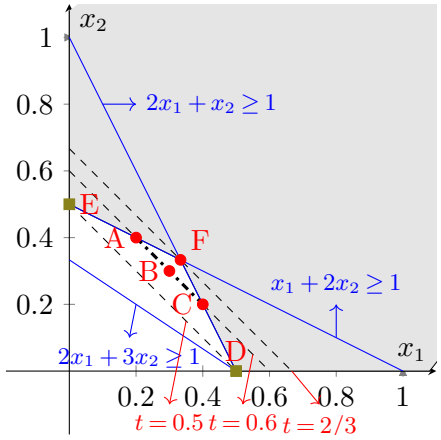

Figure 3 Illustration of ALSO-X, CVaR Approximation, and ALSO-X+ Algorithm using Example 3. Point D and point E, marked by solid square, denote the optimal solutions of the CCP, one of which is found by the ALSO-X+ algorithm. Point F shows the solution found by ALSO-X and CVaR. Three dashed lines denote objective function lines of the CCP with values equal to $t = 2/3, 0.6, 0.5$ (from top to bottom), respectively. In ALSO-X, when $t = 0.6$, point A, point B, and point C are three distinct optimal solutions.

Appendix D. The Closedness of the Feasible Region of Chance Constraint in CCP (1)

Proposition 11 *Suppose set $\mathcal{X} \subseteq \mathbb{R}^n$ is closed and function $g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ is lower semi-continuous with respect to $\boldsymbol{x}$ with probability 1, then the feasible region of CCP* (1)

$$\mathcal{X}_1 = \left\{ \boldsymbol{x} \in \mathcal{X} \colon \mathbb{P}\left\{ g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0 \right\} \geq 1 - \varepsilon \right\}$$

is closed.

Proof. For any sequence $\{\boldsymbol{x}_i\} \in \mathcal{X}_1$ converging to $\boldsymbol{x}_0$, we want to prove that $\boldsymbol{x}_0 \in \mathcal{X}_1$. Since $\mathbb{P}\{g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0\} = \mathbb{E}[\mathbb{I}(g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0)]$, then we can write set $\mathcal{X}_1$ as

$$\mathcal{X}_1 = \left\{ \boldsymbol{x} \in \mathcal{X} \colon \mathbb{E}[\mathbb{I}(g(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \leq 0)] \geq 1 - \varepsilon \right\}.$$

Since the sequence $\{\boldsymbol{x}_i\} \subseteq \mathcal{X}_1$, we have

$$\limsup_{i \to \infty} \mathbb{E}[\mathbb{I}(g(\boldsymbol{x}_i, \tilde{\boldsymbol{\xi}}) \leq 0)] \geq 1 - \varepsilon.$$

According to Fatou's lemma (see, e.g., section 4 in Royden and Fitzpatrick 1988), we have

$$\mathbb{E}\left[\limsup_{i \to \infty} \mathbb{I}(g(\boldsymbol{x}_i, \tilde{\boldsymbol{\xi}}) \leq 0) \right] \geq \limsup_{i \to \infty} \mathbb{E}[\mathbb{I}(g(\boldsymbol{x}_i, \tilde{\boldsymbol{\xi}}) \leq 0)].$$

Since the indicator function is upper semi-continuous, we have

$$\mathbb{E}\left[\mathbb{I}(\limsup_{i \to \infty} g(\boldsymbol{x}_i, \tilde{\boldsymbol{\xi}}) \leq 0) \right] \geq \mathbb{E}[\limsup_{i \to \infty} \mathbb{I}(g(\boldsymbol{x}_i, \tilde{\boldsymbol{\xi}}) \leq 0)].$$

Since the indicator function is nonincreasing and the fact that $\limsup_{i \to \infty} g(\cdot, \cdot) \geq \liminf_{i \to \infty} g(\cdot, \cdot)$, we have

$$\mathbb{E}[\mathbb{I}(\liminf_{i \to \infty} g(\boldsymbol{x}_i, \tilde{\boldsymbol{\xi}}) \leq 0)] \geq \mathbb{E}[\mathbb{I}(\limsup_{i \to \infty} g(\boldsymbol{x}_i, \tilde{\boldsymbol{\xi}}) \leq 0)].$$

According to the assumption that function $g(\boldsymbol{x}, \boldsymbol{\xi})$ is lower semi-continuous and the fact that the indicator function is nonincreasing, we have

$$\mathbb{E}[\mathbb{I}(g(\boldsymbol{x}_0, \tilde{\boldsymbol{\xi}}) \leq 0)] \geq \mathbb{E}[\mathbb{I}(\liminf_{i \to \infty} g(\boldsymbol{x}_i, \tilde{\boldsymbol{\xi}}) \leq 0)],$$

which implies that $\mathbb{E}[\mathbb{I}(g(\boldsymbol{x}_0, \tilde{\boldsymbol{\xi}}) \leq 0)] \geq \limsup_{i \to \infty} \mathbb{E}[\mathbb{I}(g(\boldsymbol{x}_i, \tilde{\boldsymbol{\xi}}) \leq 0)] \geq 1 - \varepsilon$. Thus, $\boldsymbol{x}_0 \in \mathcal{X}_1$, which completes the proof. □

We remark that Proposition 11 generalizes proposition 1.7. of Kall et al. 1994, where the authors showed that when function $g(\boldsymbol{x}, \boldsymbol{\xi})$ is continuous, the feasible region of CCP (1) is closed.

Appendix E. Tractability of ALSO-X Under Discrete Support or Elliptical Distributions

E.1 Tractability of ALSO-X Under Discrete Support

If the underlying probability distribution is finite-support with N scenarios, i.e., the random vector $\tilde{\boldsymbol{\xi}}$ has a finite support $\Xi = \{\boldsymbol{\xi}^1, \cdots, \boldsymbol{\xi}^N\}$ with $\mathbb{P}\{\tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}^i\} = p_i$ for all $i \in [N]$, then CCP (1) reduces to

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \sum_{i\in[N]} p_i \mathbb{I}(g(\boldsymbol{x}, \boldsymbol{\xi}^i) \le 0) \ge 1-\varepsilon \right\}, \tag{44}$$

and by projecting out functional variable $s(\cdot)$, ALSO-X (7) admits the following form

$$\begin{aligned} v^A = \min_{t} \quad & t, \\ \text{s.t.} \quad & \boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \sum_{i\in[N]} p_i [g(\boldsymbol{x}, \boldsymbol{\xi}^i)]_+ \colon \boldsymbol{c}^\top \boldsymbol{x} \le t \right\}, \\ & \sum_{i\in[N]} p_i \mathbb{I}(g(\boldsymbol{x}^*, \boldsymbol{\xi}^i) \le 0) \ge 1-\varepsilon. \end{aligned} \tag{45}$$

As a direct application of theorem A.3.3. in Ben-Tal et al. 2009, the following corollary shows that under mild conditions, the hinge-loss approximation (45) can be tractable.

Corollary 3 *(theorem A.3.3. in Ben-Tal et al. 2009) Suppose that (i) the encoding length of t is polynomial in that of CCP* (44)*; and (ii) the feasible region of the hinge-loss approximation is contained in a Euclidean ball with radius R and is containing a Euclidean ball with radius r. Then there exists an efficient algorithm to solve the hinge-loss approximation* (45) *to $\widehat{\varepsilon} > 0$ accuracy, whose running time is polynomial in $n, m, I, N, \ln(R/r), \ln(1/\widehat{\varepsilon})$, and the encoding length of CCP* (44)*.*

E.2 Tractability of ALSO-X Under Elliptical Distributions

For the single linear CCP (1), i.e., $I = 1$ and $g(\boldsymbol{x}, \boldsymbol{\xi}) = \boldsymbol{\xi}^\top \boldsymbol{a}_1(\boldsymbol{x}) - b_1(\boldsymbol{x})$ with affine functions $\boldsymbol{a}_1(\boldsymbol{x})$, $b_1(\boldsymbol{x})$, if the random parameters $\tilde{\boldsymbol{\xi}}$ follow a joint elliptical distribution with $\tilde{\boldsymbol{\xi}} \sim \mathbb{P}_{\mathrm{E}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \widehat{g})$, CCP (1) reduces to

$$v^* = \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon b_1(\boldsymbol{x}) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x}) \ge \Phi^{-1}(1-\varepsilon) \sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top \boldsymbol{\Sigma} \boldsymbol{a}_1(\boldsymbol{x})} \right\}, \tag{46}$$

and by projecting out variable α in (10), ALSO-X admits the following form

$$v^A = \min_{t} \quad t$$

$$\begin{aligned}\text{s.t.}\quad \boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathcal{X}} \Bigg\{ & \left(1-\Phi\left(\frac{b_1(\boldsymbol{x})-\boldsymbol{\mu}^\top\boldsymbol{x}}{\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}}\right)\right)\left(\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x})-b_1(\boldsymbol{x})\right) \\ & + \sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}\overline{G}\left(\frac{1}{2}\left(\frac{b_1(\boldsymbol{x})-\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x})}{\sqrt{\boldsymbol{a}_1(\boldsymbol{x})^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x})}}\right)^2\right) : \boldsymbol{c}^\top\boldsymbol{x}\le t\Bigg\}, \\ & b_1(\boldsymbol{x}^*)-\boldsymbol{\mu}^\top\boldsymbol{a}_1(\boldsymbol{x}^*) \ge \Phi^{-1}(1-\varepsilon)\sqrt{\boldsymbol{a}_1(\boldsymbol{x}^*)^\top\boldsymbol{\Sigma}\boldsymbol{a}_1(\boldsymbol{x}^*)}.\end{aligned} \tag{47}$$

Similarly, the following corollary shows that under mild conditions, the hinge-loss approximation (47) can be tractable.

Corollary 4 *(theorem A.3.3. in Ben-Tal et al. 2009) Suppose that (i) the encoding length of t is polynomial in that of CCP* (46)*; and (ii) the feasible region of the hinge-loss approximation is contained in a Euclidean ball with radius R and is containing a Euclidean ball with radius r. Then there exists an efficient algorithm to solve the hinge-loss approximation* (47) *to $\widehat{\varepsilon}>0$ accuracy, whose running time is polynomial in $n, m, \ln(R/r), \ln(1/\widehat{\varepsilon})$, and the encoding length of CCP* (46)*.*

Appendix F. An Example when ALSO-X Fails to Find any Feasible Solution

Example 11 Consider a CCP with 3 equiprobable scenarios (i.e., $N=3$, $\mathbb{P}\{\tilde{\boldsymbol{\xi}}=\boldsymbol{\xi}^i\}=1/N$), risk level $\varepsilon=1/3$, set $\mathcal{X}=\mathbb{R}^2_+$, function $g(\boldsymbol{x},\boldsymbol{\xi})=-\boldsymbol{\xi}_1^\top\boldsymbol{x}+\xi_2$, and $\boldsymbol{\xi}_1^1=(1,0)^\top$, $\boldsymbol{\xi}_1^2=(0,1)^\top$, $\boldsymbol{\xi}_1^3=(1,1)^\top$, $\xi_2^1=\xi_2^2=1,\xi_2^3=-1$. The optimal value of this CCP can be found by solving the following mixed-integer linear program

$$v^*=\min_{\boldsymbol{x}\in\mathbb{R}^2_+}\left\{x_1+x_2\colon \mathbb{I}(x_1\geq 1)+\mathbb{I}(x_2\geq 1)+\mathbb{I}(-x_1-x_2\geq -1)\geq 2\right\},$$

i.e., $v^*=1$.

ALSO-X of this example may be infeasible, which can be formulated as

$$\begin{aligned}v^A=\min_t\Bigg\{t\colon\ &\sum_{i\in[3]}\mathbb{I}(s_i^*=0)\geq 2,\\ (\boldsymbol{x}^*,\boldsymbol{s}^*)\in\ &\operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^2_+,\boldsymbol{s}\in\mathbb{R}^3_+}\left\{\frac{1}{3}\sum_{i\in[3]}s_i\colon x_1\geq 1-s_1,x_2\geq 1-s_2,-x_1-x_2\geq -1-s_3,x_1+x_2\leq t\right\}\Bigg\}.\end{aligned}$$

When $t\geq 1$, the hinge-loss approximation returns a solution with $x_1^*=1/2,x_2^*=1/2,s_1^*=1/2,s_2^*=1/2,s_3^*=0$, and the support size of $\boldsymbol{s}^*$ is greater than 1, then we have to increase the objective bound t to the infinity. Therefore, in this example, ALSO-X cannot return any feasible solution. Simple calculations show that CVaR approximation is also infeasible in this example. ◇

Appendix G. Complexity of CCP (9) when $\varepsilon \in (0.5, 1)$

Proposition 12 *When $\varepsilon \in (0.5, 1)$, CCP (9) in general is NP-hard.*

Proof. Let us first consider the NP-hard problem - optimization of a general binary program (Garey 1979), which asks

Optimization of a general binary program. *Given an integer matrix $\boldsymbol{D} \in \mathbb{Z}^{m\times n}$, and integer vector $\boldsymbol{d} \in \mathbb{Z}^m$, what is an optimal solution of the problem $\min_{\boldsymbol{x}\in\{0,1\}^n}\{\boldsymbol{c}^\top \boldsymbol{x} : \boldsymbol{D}\boldsymbol{x} \geq \boldsymbol{d}\}$?*

Consider a special case of CCP (9), where set $\mathcal{X} = \{(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{D}\boldsymbol{x} \geq \boldsymbol{d}, \boldsymbol{x} + \boldsymbol{y} = \boldsymbol{e}, \boldsymbol{x}, \boldsymbol{y} \in [0,1]^n\}$, affine functions $b_1(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{\mu}^\top \boldsymbol{a}_1(\boldsymbol{x}, \boldsymbol{y}) = \Phi^{-1}(1-\varepsilon)\sqrt{n}$ and $\boldsymbol{a}_1(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}, \boldsymbol{y})$, the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{I}_{2n}$. In this case, CCP (9) can be rewritten as

$$v^* = \min_{\boldsymbol{x},\boldsymbol{y}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \Phi^{-1}(1-\varepsilon)\sqrt{n} \geq \Phi^{-1}(1-\varepsilon) \sqrt{\sum_{i\in[n]} (x_i^2 + y_i^2)}, \boldsymbol{x} + \boldsymbol{y} = \boldsymbol{e}, \boldsymbol{D}\boldsymbol{x} \geq \boldsymbol{d}, \boldsymbol{x}, \boldsymbol{y} \in [0,1]^n \right\}.$$

Since $\varepsilon \in (0.5, 1)$, we must have $\Phi^{-1}(1-\varepsilon) < 0$ and CCP (9) is

$$v^* = \min_{\boldsymbol{x},\boldsymbol{y}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon n \leq \sum_{i\in[n]} (x_i^2 + y_i^2), \boldsymbol{x} + \boldsymbol{y} = \boldsymbol{e}, \boldsymbol{D}\boldsymbol{x} \geq \boldsymbol{d}, \boldsymbol{x}, \boldsymbol{y} \in [0,1]^n \right\}.$$

Since for each $i \in [n]$, the maximization problem $\max_{x_i, y_i \in [0,1]} (x_i^2 + y_i^2) = 1$ has two optimal solutions $x_i = 0, y_i = 1$ or $x_i = 1, y_i = 0$, the constraint $n \leq \sum_{i\in[n]} (x_i^2 + y_i^2)$ is satisfied if and only if $\boldsymbol{x} \in \{0,1\}^n$ and $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{e}$. Thus, projecting out variables $\boldsymbol{y}$, CCP (9) can be further reduced to

$$v^* = \min_{\boldsymbol{x}} \left\{ \boldsymbol{c}^\top \boldsymbol{x} \colon \boldsymbol{D}\boldsymbol{x} \geq \boldsymbol{d}, \boldsymbol{x} \in \{0,1\}^n \right\}.$$

which is exactly the desirable binary program. This completes the proof. □

Appendix H. Comparing ALSO-X+ Algorithm 4 and Exact Big-M Model

Big-M model is known to work well for solving a CCP (Ahmed et al. 2017). Albeit being a heuristic, the proposed ALSO-X+ Algorithm 4 can effectively identify better feasible solutions than the exact Big-M model with a much shorter solution time. To illustrate this, we use "UB" and "LB" to denote the best upper bound and the best lower bound found by the Big-M model. Since we may not be able to solve the Big-M model to optimality within the time limit, we use GAP to denote its optimality gap as

$$\text{GAP}(\%) = \frac{|\text{UB} - \text{LB}|}{|\text{LB}|} \times 100,$$

while we use the term "Improvement" to denote the solution quality of ALSO-X+ Algorithm 4

$$\text{Improvement}(\%) = \frac{\text{UB} - \text{value of the ALSO-X+ Algorithm 4}}{|\text{UB}|} \times 100.$$

The numerical results are shown in Table 4 and Table 5. It is seen that for most instances, especially for those with a larger problem dimension, the Big-M model cannot be solved to optimality, while ALSO-X+ Algorithm 4 can provide better solutions than the best upper bounds found by the Big-M model in a much shorter time, and ALSO-X+ can consistently find near-optimal solutions or even optimal solutions, which further validates the efficacy of our proposed methods.

Table 4 Comparisons Between the Exact Big-Model and ALSO-X+ Algorithm 4 for Solving the Nonlinear CCP with Small Instances

N	n	$\varepsilon = 0.05$				$\varepsilon = 0.10$			
		Big-M Model		ALSO-X+		Big-M Model		ALSO-X+	
		Gap (%)	Time (s)	Improvement (%)	Time (s)	Gap (%)	Time (s)	Improvement (%)	Time (s)
30	20	0.00	5.87	0.00	5.58	0.00	8.93	0.00	7.31
	40	0.00	12.62	0.00	8.46	0.00	19.53	0.00	9.89
	100	0.00	2076.30	-0.21	8.98	0.00	3454.75	-0.22	10.03
40	20	0.00	7.68	0.00	8.67	0.00	12.53	0.00	11.37
	40	0.00	23.45	0.00	11.97	0.00	263.56	0.00	15.84
	100	2.44	3600	-0.30	17.39	6.86	3600	-0.92	20.53
50	20	0.00	15.22	-0.23	10.87	0.00	98.01	-0.33	11.93
	40	0.00	66.36	-0.32	12.46	0.00	2190.43	-0.18	13.71
	100	3.42	3600	-0.81	18.53	19.67	3600	0.05	25.49

Table 5 Comparisons Between the Exact Big-Model and ALSO-X+ Algorithm 4 for Solving the Nonlinear CCP with Large Instances

N	n	$\varepsilon = 0.05$				$\varepsilon = 0.10$			
		Big-M Model		ALSO-X+		Big-M Model		ALSO-X+	
		Gap (%)	Time (s)	Improvement (%)	Time (s)	Gap (%)	Time (s)	Improvement (%)	Time (s)
400	20	14.48	3600	1.31	16.74	14.83	3600	0.47	13.29
	40	16.72	3600	1.93	16.84	29.81	3600	1.99	16.53
	100	198.71	3600	2.81	31.68	58.20	3600	2.04	25.25
600	20	12.37	3600	1.95	17.61	27.51	3600	2.19	22.08
	40	25.74	3600	1.86	18.39	35.14	3600	1.69	22.88
	100	251.84	3600	1.33	28.21	78.06	3600	2.63	31.60
1000	20	17.70	3600	2.55	20.24	27.76	3600	1.52	34.48
	40	31.85	3600	3.11	21.74	55.97	3600	2.87	47.55
	100	421.88	3600	2.18	33.50	345.89	3600	2.71	63.30