Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Conic Programming Reformulations of Two-Stage Distributionally Robust Linear Programs over Wasserstein Balls

Grani A. Hanasusanto, Daniel Kuhn

To cite this article:
Grani A. Hanasusanto, Daniel Kuhn (2018) Conic Programming Reformulations of Two-Stage Distributionally Robust Linear Programs over Wasserstein Balls. Operations Research 66(3):849-869. https://doi.org/10.1287/opre.2017.1698

http://pubsonline.informs.org/journal/opre/

OPERATIONS RESEARCH
Vol. 66, No. 3, May–June 2018, pp. 849–869
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Conic Programming Reformulations of Two-Stage Distributionally Robust Linear Programs over Wasserstein Balls

Grani A. Hanasusanto,[a] Daniel Kuhn[b]

[a] Graduate Program in Operations Research and Industrial Engineering, University of Texas at Austin, Austin, Texas 78712;
[b] École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

Contact: grani.hanasusanto@utexas.edu, http://orcid.org/0000-0003-4900-2958 (GAH); daniel.kuhn@epfl.ch, http://orcid.org/0000-0003-2697-8886 (DK)

Received: September 26, 2016
Revised: May 10, 2017
Accepted: October 5, 2017
Published Online in Articles in Advance: May 25, 2018

Subject Classifications: stochastic programming; decision analysis: risk; probability
Area of Review: Optimization

https://doi.org/10.1287/opre.2017.1698

Abstract. Adaptive robust optimization problems are usually solved *approximately* by restricting the adaptive decisions to simple parametric decision rules. However, the corresponding approximation error can be substantial. In this paper we show that two-stage robust and distributionally robust linear programs can often be reformulated *exactly* as conic programs that scale polynomially with the problem dimensions. Specifically, when the ambiguity set constitutes a 2-Wasserstein ball centered at a discrete distribution, the distributionally robust linear program is equivalent to a copositive program (if the problem has complete recourse) or can be approximated arbitrarily closely by a sequence of copositive programs (if the problem has sufficiently expensive recourse). These results directly extend to the classical robust setting and motivate strong tractable approximations of two-stage problems based on semidefinite approximations of the copositive cone. We also demonstrate that the two-stage distributionally robust optimization problem is equivalent to a tractable linear program when the ambiguity set constitutes a 1-Wasserstein ball centered at a discrete distribution and there are no support constraints.

Funding: This research was supported by the National Science Foundation [NSF Grant 1752125] and the Swiss National Science Foundation [Grant BSCGI0_157733].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2017.1698.

Keywords: two-stage decision problems • distributionally robust optimization • copositive programming

1. Introduction

In two-stage optimization under uncertainty, an agent selects a here-and-now decision before observing the realization of some decision-relevant random vector. Once the uncertainty has been revealed, a wait-and-see decision is taken to correct any undesired effects of the here-and-now decision in the realized scenario. Classical stochastic programming seeks a single here-and-now decision and a family of (possibly infinitely many) wait-and-see decisions—one for each possible uncertainty realization—with the goal to minimize the sum of a deterministic here-and-now cost and the expectation of an uncertain wait-and-see cost (Shapiro et al. 2014). Classical robust optimization, by contrast, seeks decisions that minimize the worst case of the total cost across all possible uncertainty realizations (Ben-Tal et al. 2009). While stochastic programming assumes full knowledge of the distribution governing the uncertain problem parameters, which is needed to evaluate the expectation of the total costs, robust optimization denies (or ignores) any knowledge of this distribution except for its support.

Distributionally robust optimization is an alternative modeling paradigm pioneered in Dupačová (1966), Scarf (1958), Shapiro and Kleywegt (2002). It has gained new thrust over the last decade and challenges the black-and-white view of stochastic and robust optimization. Specifically, it assumes that the decision maker has access to some limited probabilistic information (e.g., in the form of the distribution's moments, its structural properties, or its distance to a reference distribution) but not enough to precisely pin down the true distribution. In this setting, a meaningful objective is to minimize the worst-case expected total cost, where the worst case is evaluated across an ambiguity set that contains all distributions consistent with the available probabilistic information. Distributionally robust models enjoy strong theoretical justification from decision theory (Gilboa and Schmeidler 1989), and there is growing evidence that they provide high-quality decisions at a moderate computational cost (Delage and Ye 2010, Goh and Sim 2010, Wiesemann et al. 2014).

Two-stage decision problems under uncertainty—whether stochastic, robust, or distributionally robust—typically involve a continuum of wait-and-see decisions and thus constitute infinite-dimensional functional optimization problems. Therefore, they can only be solved approximately, except in contrived circumstances. The existing approximation methods can roughly be subdivided into *discretization schemes* (Hadjiyiannis et al. 2011, Kleywegt et al. 2002, Shapiro 2003) and *decision rule methods* (Ben-Tal et al. 2004, Georghiou et al. 2015,

Goh and Sim 2010). Discretization schemes approximate the support of the uncertain parameters with a finite subset, which entails a relaxation of the original problem and encourages optimistically biased solutions. Decision rule methods, on the other hand, approximate the infinite-dimensional space of all wait-and-see decisions with a finite-dimensional subspace of linearly parameterized decision rules, which entails a restriction of the original problem and leads to pessimistically biased solutions. In this paper, we introduce a new method for approximating two-stage distributionally robust linear programs, which can be classified neither as a discretization scheme nor as a decision rule method. We first reformulate the original infinite-dimensional optimization problem as an equivalent finite-dimensional conic program of polynomial size, which absorbs all the complexity in its cones, and then we replace the cones with tractable inner approximations.

Our exposition focuses on distributionally robust linear programs whose ambiguity sets contain all discrete and continuous distributions supported on a polytope that have a Wasserstein distance of at most ϵ from a discrete reference distribution (such as the empirical distribution corresponding to finitely many samples from the unknown true distribution). This problem class encapsulates the two-stage stochastic linear programs with discrete distributions (for $\epsilon = 0$) and the two-stage robust optimization problems with bounded polyhedral uncertainty sets (for $\epsilon = \infty$) as special cases. Wasserstein ambiguity sets were first used in the context of portfolio optimization (Pflug and Wozabal 2007). The corresponding distributionally robust optimization models were initially perceived as difficult and thus tackled with methods from global optimization (Wozabal 2012); see also Pflug and Pichler (2014, chap. 7). Recently, it has been discovered, however, that distributionally robust optimization problems with Wasserstein ambiguity sets can often be reformulated as finite convex programs (Mohajerin Esfahani and Kuhn 2017, Zhao and Guan 2018). Single-stage problems with piecewise linear cost functions, for instance, are tractable and admit convex reformulations of polynomial sizes (Mohajerin Esfahani and Kuhn 2017). Two-stage problems, on the other hand, are generically NP-hard. Their convex reformulations have exponential size but are amenable to Benders-type decomposition algorithms (Zhao and Guan 2018). Alternatively, two-stage problems can be converted to single-stage problems via a decision rule approximation, in which case they admit again a convex reformulation of polynomial size and thus regain tractability (Gao and Kleywegt 2016).

This paper extends the state of the art in two-stage distributionally robust linear programming along several dimensions. We highlight the following main contributions:

(i) We prove that any two-stage distributionally robust linear program with *complete recourse* is equivalent to a copositive program of polynomial size if the ambiguity set constitutes a 2-Wasserstein ball centered at a discrete distribution.

(ii) We prove that any two-stage distributionally robust linear program with *sufficiently expensive recourse* can be approximated arbitrarily closely by a sequence of copositive programs of a fixed polynomial size if the ambiguity set constitutes a 2-Wasserstein ball centered at a discrete distribution.

(iii) By using nested hierarchies of tractable convex cones to approximate the (intractable) copositive cones from the inside (Bomze and de Klerk 2002, de Klerk and Pasechnik 2002, Parrilo 2000), we obtain sequences of tractable conservative approximations for the two-stage distributionally robust linear programs described in items i and ii. These approximations can be made arbitrarily accurate. However, numerical tests suggest that even the coarsest of these approximations distinctly outperform the state-of-the-art decision rule approximations in terms of accuracy.

(iv) We prove that any two-stage distributionally robust linear program with *fixed costs* is equivalent to a tractable linear program if the ambiguity set constitutes a 1-Wasserstein ball centered at a discrete distribution and if there are no support constraints. We also show that this tractability result is sharp.

(v) We demonstrate that all of the above results carry directly over to classical two-stage robust optimization problems with bounded polyhedral uncertainty sets. To our best knowledge, we provide the first (polynomially sized) conic programming reformulations for generic problem instances in this class.

Two-stage distributionally robust linear programs with objective uncertainty are studied in Bertsimas et al. (2010). Assuming that only the first- and second-order moments of the uncertain cost coefficients are known, these problems can be reformulated as tractable semidefinite programs. In the presence of constraint uncertainty, however, these problems become intractable. Two-stage distributionally robust binary programs with polyhedral moment information are studied in Hanasusanto et al. (2016b). If only the cost coefficients are uncertain, these problems can be reformulated as explicit mixed-integer linear programs of polynomial sizes. While two-stage distributionally robust optimization endeavors to minimize the worst-case (maximal) expected wait-and-see cost, a parallel stream of research investigates the best-case (minimal) expectations of the minima of mixed zero-one linear programs with objective uncertainty. Under first- and second-order moment information, any such best-case expectation can be reformulated as the optimal value of a completely positive program (Natarajan

et al. 2011). In fact, this best-case expectation even reduces to the optimal value of a tractable semidefinite program whenever the convex hull of all rank-1 outer products of feasible wait-and-see decisions with themselves is semidefinite-representable (Natarajan and Teo 2017). These deep theoretical results have recently opened up new avenues for modeling and solving stochastic appointment scheduling problems (Kong et al. 2013) and also have ramifications for computing best-worst choice probabilities in discrete choice models (Natarajan and Teo 2017). A comprehensive survey of recent results at the interface of distributionally robust optimization and completely positive programming is provided in Li et al. (2014).

In contrast to the existing literature, here we develop copositive programming reformulations for generic two-stage distributionally robust linear programs where both the objective function and the constraints may be affected by the uncertainty. We also present new linear programming reformulations for two-stage distributionally robust linear programs where the uncertainty affects only the constraints. These exact reformulations are reminiscent of the conservative approximation models for two-stage robust optimization models derived in Ardestani-Jaafari and Delage (2016) by leveraging popular reformulation-linearization techniques from bilinear programming. Another main difference to the existing literature is our focus on Wasserstein balls instead of moment ambiguity sets to capture distributional uncertainty. This has the advantage that the degree of ambiguity aversion can be controlled by tuning the radius of the Wasserstein ball.

A key benefit of Wasserstein balls is that they provide natural confidence sets for the unknown distribution of the uncertain problem parameters. Specifically, the Wasserstein ball around the empirical distribution on I independent historical samples contains the unknown true distribution with confidence $1-\beta$ if its radius exceeds an explicit threshold $\epsilon_I(\beta)$ that is known in closed form (Mohajerin Esfahani and Kuhn 2017, Zhao and Guan 2018). Thus, the corresponding distributionally robust optimization problem offers a $1-\beta$ upper confidence bound on the optimal value of the true stochastic program. One can also show that this data-driven distributionally robust optimization problem converges to the corresponding true stochastic program as the sample size I tends to infinity (Mohajerin Esfahani and Kuhn 2017, Zhao and Guan 2018). Other data-driven distributionally robust optimization models that offer finite sample and asymptotic guarantees are discussed in Bertsimas et al. (2017) based on goodness-of-fit ambiguity sets, in Jiang and Guan (2018) based on L^1-norm ball ambiguity sets, and in Love and Bayraksan (2016) based on Φ-divergence ambiguity sets.

While this paper was under review, we became aware of the paper by Xu and Burer (Xu and Burer 2018), which was submitted simultaneously. It turns out that our Corollary 1 is equivalent to theorem 1 in Xu and Burer (2018), and so we mention it here for the readers' reference. While Xu and Burer (2018) focus on two-stage *robust* linear programs with right-hand-side uncertainty, we develop copositive programming reformulations for *distributionally robust* two-stage linear programs with objective *and* constraint uncertainty.

The rest of this paper is structured as follows. Section 2 provides a formal problem statement and reviews some fundamental results from Mohajerin Esfahani and Kuhn (2017) and Zhao and Guan (2018). In Section 3 we derive copositive programming reformulations for two-stage distributionally robust linear programs over 2-Wasserstein balls and discuss tractable approximations. Exact tractable linear programming reformulations for two-stage distributionally robust linear programs over 1-Wasserstein balls are described in Section 4. Section 5 reports on numerical results.

1.1. Notation

For any $I \in \mathbb{N}$, we define $[I]$ as the index set $\{1,\ldots,I\}$. We denote by $\mathbb{I}$ the identity matrix and by $\mathbf{e}$ the vector of all ones. Their dimensions will be clear from the context. The trace of a square matrix $\mathbf{M}$ is denoted as $\mathrm{tr}(\mathbf{M})$. We define $\mathrm{diag}(\mathbf{v})$ as the diagonal matrix with the vector $\mathbf{v}$ on its main diagonal. The set of nonnegative (positive) reals is denoted as $\mathbb{R}_+$ ($\mathbb{R}_{++}$). The set of all symmetric matrices in $\mathbb{R}^{K\times K}$ is denoted as $\mathbb{S}^K$, while the cone of positive semidefinite matrices in $\mathbb{R}^{K\times K}$ is denoted as $\mathbb{S}^K_+$. We define the cone of copositive matrices as $\mathscr{C} = \{\mathbf{M}\in\mathbb{S}^K : \boldsymbol{\xi}^\top\mathbf{M}\boldsymbol{\xi}\geqslant 0\ \forall\,\boldsymbol{\xi}\geqslant\mathbf{0}\}$ and the cone of completely positive matrices as $\mathscr{C}^* = \{\mathbf{M}\in\mathbb{S}^K : \mathbf{M}=\mathbf{B}\mathbf{B}^\top \text{ for some } \mathbf{B}\in\mathbb{R}_+^{K\times g(K)}\}$, where $g(K) = \max\left\{\binom{K+1}{2}-4, K\right\}$ (Shaked-Monderer et al. 2015). For any $\mathbf{Q},\mathbf{R}\in\mathbb{S}^K$, the relations $\mathbf{Q}\geq\mathbf{R}$, $\mathbf{Q}\geq_{\mathscr{C}}\mathbf{R}$, and $\mathbf{Q}\geq_{\mathscr{C}^*}\mathbf{R}$ mean that $\mathbf{Q}-\mathbf{R}$ is an element of $\mathbb{S}^K_+$, $\mathscr{C}$, and $\mathscr{C}^*$, respectively. We denote the jth row (jth column) of a matrix $\mathbf{M}$ as $\mathbf{M}_{j:}$ ($\mathbf{M}_{:j}$). All random variables are designated by tildes (e.g., $\tilde{\boldsymbol{\xi}}$), while their realizations are denoted without tildes (e.g., $\boldsymbol{\xi}$). The characteristic function of a set $\mathscr{S}$ is defined as $\chi_{\mathscr{S}}(\boldsymbol{\xi}) = 0$ if $\boldsymbol{\xi}\in\mathscr{S}$ and as $=\infty$ otherwise.

2. Problem Formulation

We study two-stage distributionally robust linear programs of the form

$$\begin{array}{ll}\text{minimize} & \mathbf{c}^\top\mathbf{x} + \mathscr{Z}(\mathbf{x})\\ \text{subject to} & \mathbf{x}\in\mathscr{X},\end{array} \tag{1}$$

where $\mathscr{X}\subseteq\mathbb{R}^{N_1}$ is the feasible set of the here-and-now decisions, $\mathbf{c}^\top\mathbf{x}$ is the here-and-now cost, and $\mathscr{Z}(\mathbf{x})$ is the worst-case expected wait-and-see cost. Formally, we set

$$\mathscr{Z}(\mathbf{x}) = \sup_{\mathbb{P}\in\hat{\mathscr{P}}} \mathbb{E}_{\mathbb{P}}[Z(\mathbf{x},\tilde{\boldsymbol{\xi}})], \tag{2}$$

where $\tilde{\boldsymbol{\xi}} \in \Xi \subseteq \mathbb{R}^K$ is a random vector comprising the uncertain problem parameters and $\hat{\mathcal{P}}$ is an ambiguity set that contains the possible distributions of $\tilde{\boldsymbol{\xi}}$. The recourse function $Z(\mathbf{x}, \boldsymbol{\xi})$ in (2) constitutes the optimal value of the recourse problem; that is,

$$\begin{aligned} Z(\mathbf{x}, \boldsymbol{\xi}) = \inf\ & (\mathbf{Q}\boldsymbol{\xi} + \mathbf{q})^\top \mathbf{y} \\ \text{s.t. } & \mathbf{y} \in \mathbb{R}^{N_2}, \\ & \mathbf{T}(\mathbf{x})\boldsymbol{\xi} + \mathbf{h}(\mathbf{x}) \leqslant \mathbf{W}\mathbf{y}, \end{aligned} \tag{3}$$

where $\mathbf{T}(\mathbf{x}) \in \mathbb{R}^{M \times K}$ and $\mathbf{h}(\mathbf{x}) \in \mathbb{R}^M$ are matrix- and vector-valued affine functions, respectively. The dual of the recourse problem is given by

$$\begin{aligned} Z_{\mathrm{d}}(\mathbf{x}, \boldsymbol{\xi}) = \sup\ & (\mathbf{T}(\mathbf{x})\boldsymbol{\xi} + \mathbf{h}(\mathbf{x}))^\top \mathbf{p} \\ \text{s.t. } & \mathbf{p} \in \mathbb{R}^M_+, \\ & \mathbf{Q}\boldsymbol{\xi} + \mathbf{q} = \mathbf{W}^\top \mathbf{p}. \end{aligned} \tag{4}$$

We introduce the following standard terminology that will be used throughout the paper.

Definition 1 (Complete Recourse)**.** We say that the two-stage distributionally robust linear program (1) has complete recourse if there exists $\mathbf{y}^+ \in \mathbb{R}^{N_2}$ with $\mathbf{W}\mathbf{y}^+ > \mathbf{0}$.

Complete recourse implies that problem (3) is feasible for every $\mathbf{x} \in \mathbb{R}^{N_1}$ and $\boldsymbol{\xi} \in \mathbb{R}^K$. Indeed, it implies that there is always a $\lambda > 0$ such that $\mathbf{y} = \lambda \mathbf{y}^+$ exceeds $\mathbf{T}(\mathbf{x})\boldsymbol{\xi} + \mathbf{h}(\mathbf{x})$.

Definition 2 (Sufficiently Expensive Recourse)**.** We say that the two-stage distributionally robust linear program (1) has sufficiently expensive recourse if for any fixed $\boldsymbol{\xi} \in \Xi$ the dual problem (4) is feasible.

If problem (1) has complete recourse, then $Z(\mathbf{x}, \boldsymbol{\xi}) < +\infty$ for every $\mathbf{x} \in \mathbb{R}^{N_1}$ and $\boldsymbol{\xi} \in \mathbb{R}^K$. On the other hand, if problem (1) has sufficiently expensive recourse, then $Z(\mathbf{x}, \boldsymbol{\xi}) > -\infty$ for every $\mathbf{x} \in \mathscr{X}$ and $\boldsymbol{\xi} \in \Xi$. If both conditions are satisfied, then $Z(\mathbf{x}, \boldsymbol{\xi})$ is finite. Each condition by itself implies that strong duality holds between the primal and dual linear programs (3) and (4), respectively. Throughout this paper, we will always assume that problem (1) has sufficiently expensive recourse. This is a weak condition that is satisfied by many problems even with induced constraints. The complete recourse assumption, which rules out induced constraints, will only be imposed occasionally to obtain stronger results.

Following Mohajerin Esfahani and Kuhn (2017) and Zhao and Guan (2018), we assume henceforth that the true distribution of $\tilde{\boldsymbol{\xi}}$ is unknown but that we have access to I samples $\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_I$ from this distribution. In this case, we can define the empirical distribution $\hat{\mathbb{P}}_I = (1/I)\sum_{i \in [I]} \delta_{\hat{\boldsymbol{\xi}}_i}$—that is, the uniform distribution on the samples. The ambiguity set $\hat{\mathcal{P}}$ in (1) can then be defined as the family of all distributions that are close to the empirical distribution $\hat{\mathbb{P}}_I$ with respect to the Wasserstein metric.

Definition 3 (Wasserstein Metric)**.** For any $r \geqslant 1$, let $\mathscr{M}^r(\Xi)$ be the set of all probability distributions $\mathbb{P}$ supported on Ξ satisfying $\mathbb{E}_{\mathbb{P}}[d(\tilde{\boldsymbol{\xi}}, \boldsymbol{\xi}_0)^r] = \int_\Xi d(\boldsymbol{\xi}, \boldsymbol{\xi}_0)^r \mathbb{P}(\mathrm{d}\boldsymbol{\xi}) < \infty$, where $\boldsymbol{\xi}_0 \in \Xi$ is some reference point and $d(\boldsymbol{\xi}, \boldsymbol{\xi}_0)$ is a continuous reference metric on Ξ. For any $r \geqslant 1$, the r-Wasserstein distance between two distributions $\mathbb{P}_1, \mathbb{P}_2 \in \mathscr{M}^r(\Xi)$ is defined as

$$W^r(\mathbb{P}_1, \mathbb{P}_2) = \inf\left\{ \left(\int_{\Xi^2} d(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)^r \, \mathbb{Q}(\mathrm{d}\boldsymbol{\xi}_1, \mathrm{d}\boldsymbol{\xi}_2) \right)^{1/r} : \begin{array}{l} \mathbb{Q} \text{ is a joint distribution of } \boldsymbol{\xi}_1 \text{ and } \boldsymbol{\xi}_2 \\ \text{with marginals } \mathbb{P}_1 \text{ and } \mathbb{P}_2, \text{ respectively} \end{array} \right\}.$$

We denote the Wasserstein ball of radius ϵ centered at the empirical distribution by

$$\mathscr{B}^r_\epsilon(\hat{\mathbb{P}}_I) = \{\mathbb{P} \in \mathscr{M}^r(\Xi) \colon W^r(\mathbb{P}, \hat{\mathbb{P}}_I) \leqslant \epsilon\}.$$

The following theorem, which is adapted from Mohajerin Esfahani and Kuhn (2017) and Gao and Kleywegt (2016) and relies on the Knothe–Rosenblatt rearrangement (Villani 2008), establishes that the worst-case expectation (2) over a Wasserstein ambiguity set $\hat{\mathcal{P}} = \mathscr{B}^r_\epsilon(\hat{\mathbb{P}}_I)$ can be reformulated in terms of a generalized moment problem and the corresponding dual robust optimization problem. To keep this paper self-contained, we prove this theorem in the online appendix.

Theorem 1. *If $\hat{\mathcal{P}} = \mathscr{B}^r_\epsilon(\hat{\mathbb{P}}_I)$, the worst-case expectation (2) coincides with the optimal value of the generalized moment problem*:

$$\begin{aligned} \mathscr{Z}(\mathbf{x}) = \sup\ & \frac{1}{I}\sum_{i \in [I]} \int_\Xi Z(\mathbf{x}, \boldsymbol{\xi}) \mathbb{P}_i(\mathrm{d}\boldsymbol{\xi}) \\ \text{s.t. } & \mathbb{P}_i \in \mathscr{M}^r(\Xi) \quad \forall i \in [I], \\ & \frac{1}{I}\sum_{i \in [I]} \int_\Xi d(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_i)^r \mathbb{P}_i(\mathrm{d}\boldsymbol{\xi}) \leqslant \epsilon^r. \end{aligned} \tag{5}$$

Furthermore, for $\epsilon > 0$, this problem admits the strong dual robust optimization problem:

$$\mathscr{Z}(\mathbf{x}) = \inf_{\lambda \in \mathbb{R}_+} \left\{ \epsilon^r \lambda + \frac{1}{I}\sum_{i \in [I]} \sup_{\boldsymbol{\xi} \in \Xi} Z(\mathbf{x}, \boldsymbol{\xi}) - \lambda d(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_i)^r \right\}. \tag{6}$$

All results of this paper directly extend to the class of two-stage *robust* optimization problems, which model the uncertainties only through their uncertainty set Ξ.

Remark 1 (Two-Stage Robust Optimization)**.** If Ξ is compact and the radius ϵ of the Wassersein ball is larger than the diameter of Ξ, then the two-stage distributionally robust linear program (1) simplifies to the two-stage robust optimization problem:

$$\begin{aligned} \text{minimize } & \left\{ \mathbf{c}^\top \mathbf{x} + \max_{\boldsymbol{\xi} \in \Xi} Z(\mathbf{x}, \boldsymbol{\xi}) \right\} \\ \text{subject to } & \mathbf{x} \in \mathscr{X}. \end{aligned} \tag{7}$$

Indeed, if we set $\epsilon \geqslant \max_{\xi,\,\xi'\in\Xi} d(\xi,\xi')$, then the Wasserstein ball $\mathbb{B}_\epsilon^r(\hat{\mathbb{P}}_I)$ contains all Dirac distributions δ_ξ, $\xi\in\Xi$. This implies that the worst-case expected cost (2) reduces to the worst-case cost $\max_{\xi\in\Xi} Z(\mathbf{x},\xi)$, irrespective of the number and positions of the samples $\hat{\xi}_i$, $i\in[I]$.

3. Copositive Programming Reformulation

Throughout this section we work with the 2-Wasserstein metric and the 2-norm reference distance $d(\xi_1,\xi_2) = \|\xi_1-\xi_2\|_2$. We further assume that the support set Ξ is a nonempty polyhedron of the form

$$\Xi = \{\xi\in\mathbb{R}_+^K : \mathbf{S}\xi \leqslant \mathbf{t}\} \tag{8}$$

for some $\mathbf{S}\in\mathbb{R}^{J\times K}$ and $\mathbf{t}\in\mathbb{R}^J$. Note that we assume, without loss of generality, that Ξ is a (possibly unbounded) subset of the nonnegative orthant. Finally, we also assume that problem (1) has sufficiently expensive recourse. We will show that under these assumptions, the two-stage distributionally robust linear program (1) admits an equivalent reformulation as a copositive program.

3.1. A Copositive Upper Bound on $\mathcal{Z}(\mathbf{x})$

To derive a copositive programming-based upper bound on $\mathcal{Z}(\mathbf{x})$, we will need the following technical lemma.

Lemma 1. *For any symmetric matrix $\mathbf{M}\in\mathbb{S}^K$, we have $\mathbf{M}\succeq_{\mathscr{C}}\mathbf{0}$ if and only if*

$$[\mathbf{z}^\top 1]\mathbf{M}[\mathbf{z}^\top 1]^\top \geqslant 0 \quad \forall\,\mathbf{z}\in\mathbb{R}_+^{K-1}. \tag{9}$$

Proof. To prove sufficiency, we recall that $\mathbf{M}\succeq_{\mathscr{C}}\mathbf{0}$ if and only if $\xi^\top\mathbf{M}\xi\geqslant 0$ for all $\xi\in\mathbb{R}_+^K$. Thus, (9) follows by focusing on those $\xi\in\mathbb{R}^K$ with $\xi_K = 1$.

To prove the converse implication, assume that (9) holds. Hence, we have

$$\begin{aligned}&[\mathbf{z}^\top 1]\mathbf{M}[\mathbf{z}^\top 1]^\top \geqslant 0 \quad \forall\,\mathbf{z}\in\mathbb{R}_+^{K-1}\\ &\quad\Longrightarrow\quad [t\mathbf{z}^\top t]\mathbf{M}[t\mathbf{z}^\top t]^\top \geqslant 0 \quad \forall\,\mathbf{z}\in\mathbb{R}_+^{K-1},\ \forall\,t\in\mathbb{R}_{++}\\ &\quad\Longrightarrow\quad [\mathbf{y}^\top t]\mathbf{M}[\mathbf{y}^\top t]^\top \geqslant 0 \quad \forall\,\mathbf{y}\in\mathbb{R}_+^{K-1},\ \forall\,t\in\mathbb{R}_{++};\end{aligned}$$

that is, for any fixed $\mathbf{y}\in\mathbb{R}_+^{K-1}$, the univariate quadratic function $[\mathbf{y}^\top t]\mathbf{M}[\mathbf{y}^\top t]^\top$ is nonnegative for all $t>0$. As this function is continuous, it must, in fact, be nonnegative for all $t\geqslant 0$. Thus, $\mathbf{M}\succeq_{\mathscr{C}}\mathbf{0}$. □

We are now ready to derive an upper bound on the worst-case expectation $\mathcal{Z}(\mathbf{x})$. This bound is expressed as the optimal value of a copositive minimization problem and can thus be used to conservatively approximate (1) with a finite-dimensional minimization problem that is principally amenable to a numerical solution.

Theorem 2 (Copositive Upper Bound). *For any fixed first-stage decision $\mathbf{x}\in\mathscr{X}$, the worst-case expectation $\mathcal{Z}(\mathbf{x})$ in (2) is bounded above by the optimal value of the copositive program:*

$$\begin{aligned}&\bar{\mathcal{Z}}(\mathbf{x})\\ &= \inf \epsilon^2\lambda + \frac{1}{I}\sum_{i\in[I]}\left[s_i + q^\top\psi_i - \lambda\|\hat{\xi}_i\|_2^2 + \sum_{j\in[N_2+J]}\phi_{ij}q_j^2\right]\\ &\quad s.t.\quad \lambda\in\mathbb{R}_+,\ s_i\in\mathbb{R},\ \psi_i,\phi_i\in\mathbb{R}^{N_2+J}\qquad \forall i\in[I]\\ &\qquad \begin{bmatrix} \lambda\mathbb{I}+\mathcal{Q}^\top\mathrm{diag}(\phi_i)\mathcal{Q} & -\frac{1}{2}\mathcal{T}(\mathbf{x})^\top-\mathcal{Q}^\top\mathrm{diag}(\phi_i)\mathcal{W}^\top & -\lambda\hat{\xi}_i-\frac{1}{2}\mathcal{Q}^\top\psi_i\\ -\frac{1}{2}\mathcal{T}(\mathbf{x})-\mathcal{W}\mathrm{diag}(\phi_i)\mathcal{Q} & \mathcal{W}\mathrm{diag}(\phi_i)\mathcal{W}^\top & \frac{1}{2}(\mathcal{W}\psi_i-\mathcal{h}(\mathbf{x}))\\ \left(-\lambda\hat{\xi}_i-\frac{1}{2}\mathcal{Q}^\top\psi_i\right)^\top & \frac{1}{2}(\mathcal{W}\psi_i-\mathcal{h}(\mathbf{x}))^\top & s_i\end{bmatrix}\succeq_{\mathscr{C}}\mathbf{0}\quad\forall i\in[I],\end{aligned} \tag{10}$$

where

$$\mathcal{Q}=\begin{bmatrix}\mathbf{Q}\\ \mathbf{S}\end{bmatrix},\quad q=\begin{bmatrix}\mathbf{q}\\ -\mathbf{t}\end{bmatrix},\quad \mathcal{T}(\mathbf{x})=\begin{bmatrix}\mathbf{T}(\mathbf{x})\\ \mathbf{0}\end{bmatrix},$$
$$\mathcal{h}(\mathbf{x})=\begin{bmatrix}\mathbf{h}(\mathbf{x})\\ \mathbf{0}\end{bmatrix},\quad\text{and}\quad \mathcal{W}=\begin{bmatrix}\mathbf{W} & \mathbf{0}\\ \mathbf{0} & -\mathbb{I}\end{bmatrix}. \tag{11}$$

Remark 2. Note that the extended recourse parameters defined in (11) combine the input data of the recourse problem (3) with the parameters characterizing the support set (8).

Proof of Theorem 2. By strong linear programming duality, which holds because problem (1) has sufficiently expensive recourse, we have $Z(\mathbf{x},\xi) = Z_\mathrm{d}(\mathbf{x},\xi)$ for every $\mathbf{x}\in\mathscr{X}$ and $\xi\in\Xi$. Recalling that $r=2$ and $d(\xi_1,\xi_2)=\|\xi_1-\xi_2\|_2$, the explicit formula (4) for the optimal value $Z_\mathrm{d}(\mathbf{x},\xi)$ of the dual recourse problem and the polyhedral representation (8) for Ξ allow us to reformulate (6) as

$$\begin{aligned}\mathcal{Z}(\mathbf{x}) &= \inf_{\lambda\geqslant 0}\left\{\epsilon^2\lambda + \frac{1}{I}\sum_{i\in[I]}\sup_{\substack{\xi\geqslant 0\\ \mathbf{S}\xi\leqslant\mathbf{t}}}\ \sup_{\substack{\mathbf{p}\geqslant\mathbf{0}\\ \mathbf{Q}\xi+\mathbf{q}=\mathbf{W}^\top\mathbf{p}}} (\mathbf{T}(\mathbf{x})\xi+\mathbf{h}(\mathbf{x}))^\top\mathbf{p} - \lambda\|\xi-\hat{\xi}_i\|_2^2\right\}\\ &= \inf_{\lambda\geqslant 0}\left\{\epsilon^2\lambda + \frac{1}{I}\sum_{i\in[I]}\sup_{\substack{\xi,\pi\geqslant 0\\ \mathcal{Q}\xi+q=\mathcal{W}^\top\pi}} (\mathcal{T}(\mathbf{x})\xi+\mathcal{h}(\mathbf{x}))^\top\pi - \lambda\|\xi-\hat{\xi}_i\|_2^2\right\},\end{aligned} \tag{12}$$

where the second equality uses the definitions in (11). Note that the first M components of the new decision variable $\pi\in\mathbb{R}^{M+J}$ correspond to the dual variable $\mathbf{p}$,

while the remaining J components represent slack variables for the support constraints $\mathbf{S}\xi \leqslant \mathbf{t}$. Next, we add the following nonconvex constraints to each of the I inner maximization problems in (12):

$$q_j^2 = (\boldsymbol{\mathcal{W}}_{:j}^\top \pi - \boldsymbol{\mathcal{Q}}_{j:}^\top \xi)^2 = (\boldsymbol{\mathcal{Q}}_{j:}^\top \xi)^2 - 2\boldsymbol{\mathcal{Q}}_{j:}^\top \xi \pi^\top \boldsymbol{\mathcal{W}}_{:j} + (\boldsymbol{\mathcal{W}}_{:j}^\top \pi)^2 \qquad \forall j \in [N_2 + J].$$

Note that these constraints are redundant, as they follow from $\boldsymbol{\mathcal{Q}}\xi + q = \boldsymbol{\mathcal{W}}^\top \pi$. As will be revealed in Section 3.2, however, these constraints ensure that the optimal value of the completely positive program dual to (10) coincides with $\mathcal{Z}(\mathbf{x})$. Thanks to a recent result from the theory of quadratic programming (Burer 2009), each of the emerging (nonconvex) quadratically constrained quadratic subproblems in (12) can be reformulated as a completely positive maximization problem. Thus, we could apply standard dualization techniques to reformulate (12) as a finite copositive minimization problem. Here, we instead pursue a more direct approach, which leverages Lemma 1. By expressing all linear and quadratic constraints of the subproblems in Lagrangian form, we can reformulate (12) as

$$\begin{aligned}
\mathcal{Z}(\mathbf{x}) = \inf_{\lambda \in \mathbb{R}_+} \Bigg\{ \epsilon^2 \lambda + \frac{1}{I} \sum_{i \in [I]} \sup_{\xi, \pi \geqslant 0} \inf_{\psi_i, \phi_i} \Bigg[& (\boldsymbol{\mathcal{T}}(\mathbf{x})\xi + \hbar(\mathbf{x}))^\top \pi \\
& - \lambda \|\xi - \hat{\xi}_i\|_2^2 + \psi_i^\top (\boldsymbol{\mathcal{Q}}\xi + q - \boldsymbol{\mathcal{W}}^\top \pi) \\
& + \sum_{j \in [N_2 + J]} \phi_{ij} (q_j^2 - (\boldsymbol{\mathcal{Q}}_{j:}^\top \xi)^2 + 2\boldsymbol{\mathcal{Q}}_{j:}^\top \xi \pi^\top \boldsymbol{\mathcal{W}}_{:j} - (\boldsymbol{\mathcal{W}}_{:j}^\top \pi)^2) \Bigg] \Bigg\} \\
\leqslant \inf_{\lambda \geqslant 0, \psi_i, \phi_i} \Bigg\{ \epsilon^2 \lambda + \frac{1}{I} \sum_{i \in [I]} \sup_{\xi, \pi \geqslant 0} \Bigg[& (\boldsymbol{\mathcal{T}}(\mathbf{x})\xi + \hbar(\mathbf{x}))^\top \pi \\
& - \lambda \|\xi - \hat{\xi}_i\|_2^2 + \psi_i^\top (\boldsymbol{\mathcal{Q}}\xi + q - \boldsymbol{\mathcal{W}}^\top \pi) \\
& + \sum_{j \in [N_2 + J]} \phi_{ij} (q_j^2 - (\boldsymbol{\mathcal{Q}}_{j:}^\top \xi)^2 + 2\boldsymbol{\mathcal{Q}}_{j:}^\top \xi \pi^\top \boldsymbol{\mathcal{W}}_{:j} - (\boldsymbol{\mathcal{W}}_{:j}^\top \pi)^2) \Bigg] \Bigg\},
\end{aligned} \tag{13}$$

where ϕ_{ij} denotes the jth entry of ϕ_i. Here, the inequality follows from interchanging the order of the supremum and the infimum operators. We observe now that the terms in square brackets constitute quadratic forms in ξ and π. Thus, by introducing auxiliary epigraphical variables s_i, $i \in [I]$, to eliminate the suprema over ξ and π, we can reformulate (13) as the quadratically parameterized semi-infinite linear program:

$$\begin{aligned}
\inf \ & \left\{ \epsilon^2 \lambda + \frac{1}{I} \sum_{i \in [I]} \left[s_i + q^\top \psi_i - \lambda \|\hat{\xi}_i\|_2^2 + \sum_{j \in [N_2 + J]} \phi_{ij} q_j^2 \right] \right\} \\
\text{s.t.} \ & \lambda \in \mathbb{R}_+, s_i \in \mathbb{R}, \psi_i, \phi_i \in \mathbb{R}^{N_2 + J} \qquad \forall i \in [I], \\
& \begin{bmatrix} \xi \\ \pi \\ 1 \end{bmatrix}^\top \begin{bmatrix} \lambda \mathbb{I} + \boldsymbol{\mathcal{Q}}^\top \operatorname{diag}(\phi_i) \boldsymbol{\mathcal{Q}} & -\frac{1}{2}\boldsymbol{\mathcal{T}}(\mathbf{x})^\top - \boldsymbol{\mathcal{Q}}^\top \operatorname{diag}(\phi_i) \boldsymbol{\mathcal{W}}^\top \\ -\frac{1}{2}\boldsymbol{\mathcal{T}}(\mathbf{x}) - \boldsymbol{\mathcal{W}} \operatorname{diag}(\phi_i) \boldsymbol{\mathcal{Q}} & \boldsymbol{\mathcal{W}} \operatorname{diag}(\phi_i) \boldsymbol{\mathcal{W}}^\top \\ (-\lambda \hat{\xi}_i - \frac{1}{2}\mathbf{Q}^\top \psi_i)^\top & \frac{1}{2}(\boldsymbol{\mathcal{W}}\psi_i - \hbar(\mathbf{x}))^\top \end{bmatrix} \\
& \qquad \begin{bmatrix} -\lambda \hat{\xi}_i - \frac{1}{2}\boldsymbol{\mathcal{Q}}^\top \psi_i \\ \frac{1}{2}(\boldsymbol{\mathcal{W}}\psi_i - \hbar(\mathbf{x})) \\ s_i \end{bmatrix} \begin{bmatrix} \xi \\ \pi \\ 1 \end{bmatrix} \geqslant 0 \qquad \forall i \in [I] \ \forall (\xi, \pi) \in \mathbb{R}_+^{K+M+J},
\end{aligned}$$

which is equivalent to the copositive program (10) by virtue of Lemma 1. □

3.2. A Completely Positive Reformulation of $\mathcal{Z}(\mathbf{x})$

We now derive the dual of the copositive program (10). As we will see later, even though (10) provides an upper bound on (2), the optimal value of its dual problem coincides with the worst-case expectation (2).

Proposition 1. *The copositive program* (10) *is dual to the following completely positive program*:

$$\begin{aligned}
\underline{\mathcal{Z}}(\mathbf{x}) = \sup \ & \left\{ \frac{1}{I} \sum_{i \in [I]} [\operatorname{tr}(\boldsymbol{\mathcal{T}}(\mathbf{x})\mathbf{Y}_i) + \hbar(\mathbf{x})^\top \gamma_i] \right\} \\
\text{s.t.} \ & \gamma_i \in \mathbb{R}_+^{M+J}, \mu_i \in \mathbb{R}_+^K, \boldsymbol{\Gamma}_i \in \mathbb{S}_+^{M+J}, \\
& \quad \boldsymbol{\Omega}_i \in \mathbb{S}_+^K, \mathbf{Y}_i \in \mathbb{R}^{K \times (M+J)} \qquad \forall i \in [I], \\
& \boldsymbol{\mathcal{Q}}\mu_i + q = \boldsymbol{\mathcal{W}}^\top \gamma_i \qquad \forall i \in [I], \\
& \boldsymbol{\mathcal{Q}}_{j:}^\top \boldsymbol{\Omega}_i \boldsymbol{\mathcal{Q}}_{j:} - 2\boldsymbol{\mathcal{Q}}_{j:}^\top \mathbf{Y}_i \boldsymbol{\mathcal{W}}_{:j} + \boldsymbol{\mathcal{W}}_{:j}^\top \boldsymbol{\Gamma}_i \boldsymbol{\mathcal{W}}_{:j} = q_j^2 \\
& \qquad \forall i \in [I] \ \forall j \in [N_2 + J], \\
& \frac{1}{I} \sum_{i \in [I]} [\operatorname{tr}(\boldsymbol{\Omega}_i) - 2\hat{\xi}_i^\top \mu_i + \hat{\xi}_i^\top \hat{\xi}_i] \leqslant \epsilon^2, \\
& \begin{bmatrix} \boldsymbol{\Omega}_i & \mathbf{Y}_i & \mu_i \\ \mathbf{Y}_i^\top & \boldsymbol{\Gamma}_i & \gamma_i \\ \mu_i^\top & \gamma_i^\top & 1 \end{bmatrix} \succeq_{\mathscr{C}^*} \mathbf{0} \qquad \forall i \in [I].
\end{aligned} \tag{14}$$

Proof. The claim follows from standard conic duality theory. Details are omitted for brevity. □

In the following, we show that the worst-case expectation $\mathcal{Z}(\mathbf{x})$ is in fact equal to the optimal value $\underline{\mathcal{Z}}(\mathbf{x})$ of the completely positive program (14).

Theorem 3 (Completely Positive Reformulation)**.** *For any fixed* $\mathbf{x} \in \mathscr{X}$, *we have* $\mathcal{Z}(\mathbf{x}) = \underline{\mathcal{Z}}(\mathbf{x})$.

Proof. Recall from Theorem 1 that $\mathcal{Z}(\mathbf{x})$ coincides with the optimal value of the moment problem (5). We first prove that $\mathcal{Z}(\mathbf{x}) \leqslant \underline{\mathcal{Z}}(\mathbf{x})$. To this end, we show that any feasible solution $\{\mathbb{P}_i\}_{i \in [I]}$ of (5) gives rise to a feasible solution to (14) with the same objective function value. Let $\mathbf{p}(\xi)$ be a measurable selector of the dual feasible set mapping $\xi \rightrightarrows \{\mathbf{p} \in \mathbb{R}_+^M : \mathbf{Q}\xi + \mathbf{q} = \mathbf{W}^\top \mathbf{p}\}$, $\xi \in \Xi$, which exists because of Rockafellar and Wets (2009, corollary 14.6) and because problem (1) has sufficiently expensive recourse. Next, define $\pi(\xi) = (\mathbf{p}(\xi), \mathbf{t} - \mathbf{S}\xi)$. By construction, we have $\boldsymbol{\mathcal{Q}}\xi + q = \boldsymbol{\mathcal{W}}^\top \pi(\xi)$ for all $\xi \in \Xi$. Next, define the following candidate solution for (14):

$$\left.\begin{aligned}
\mu_i &= \int_\Xi \xi \mathbb{P}_i(\mathrm{d}\xi), & \boldsymbol{\Omega}_i &= \int_\Xi \xi\xi^\top \mathbb{P}_i(\mathrm{d}\xi), \\
\gamma_i &= \int_\Xi \pi(\xi) \mathbb{P}_i(\mathrm{d}\xi), & \boldsymbol{\Gamma}_i &= \int_\Xi \pi(\xi)\pi(\xi)^\top \mathbb{P}_i(\mathrm{d}\xi), \\
& & \mathbf{Y}_i &= \int_\Xi \xi \pi(\xi)^\top \mathbb{P}_i(\mathrm{d}\xi)
\end{aligned}\right\} \qquad \forall i \in [I]. \tag{15}$$

Since $\tilde{\xi}$ and $\pi(\tilde{\xi})$ are nonnegative random vectors, the matrix of their moments of degree $\leqslant 2$ is completely positive. Thus, the candidate solution (15) satisfies the last constraint in (14). We further have

$$\boldsymbol{Q}\xi + q = \boldsymbol{W}^\top \pi(\xi) \quad \forall \xi \in \Xi \quad \Longrightarrow \quad \boldsymbol{Q}\mu_i + q = \boldsymbol{W}^\top \gamma_i$$

and

$$\begin{aligned}(\boldsymbol{Q}_{j:}^\top \xi + q_j)^2 = (\boldsymbol{W}_{:j}^\top \pi(\xi))^2 \quad \forall \xi \in \Xi \\ \Longrightarrow \quad \boldsymbol{Q}_{j:}^\top \boldsymbol{\Omega}_i \boldsymbol{Q}_{j:} - 2\boldsymbol{Q}_{j:}^\top \boldsymbol{\Upsilon}_i \boldsymbol{W}_{:j} + \boldsymbol{W}_{:j}^\top \boldsymbol{\Gamma}_i \boldsymbol{W}_{:j} = q_j^2\end{aligned}$$

for all $j \in [N_2 + J]$. Here, the implications follow from taking expectations with respect to $\mathbb{P}_i$ on both sides of the semi-infinite constraints. Thus, the candidate solution (15) also satisfies the first and the second constraint systems in (14). Next, the feasibility of $\{\mathbb{P}_i\}_{i\in[I]}$ in the generalized moment problem (5) implies that $(1/I)\sum_{i\in[I]} \int_\Xi \|\xi - \hat{\xi}_i\|_2^2 \, \mathbb{P}_i(\mathrm{d}\xi) \leqslant \epsilon^2$. Expanding the squared norm term, we obtain

$$\begin{aligned}\epsilon^2 &\geqslant \frac{1}{I}\sum_{i\in[I]} \int_\Xi [\xi^\top \xi - 2\hat{\xi}_i^\top \xi + \hat{\xi}_i^\top \hat{\xi}_i] \, \mathbb{P}_i(\mathrm{d}\xi) \\ &= \frac{1}{I}\sum_{i\in[I]} [\mathrm{tr}(\boldsymbol{\Omega}_i) - 2\hat{\xi}_i^\top \mu_i + \hat{\xi}_i^\top \hat{\xi}_i],\end{aligned}$$

and thus the candidate solution (15) also satisfies the penultimate constraint in (14). Finally, the objective function of (5) can be reformulated as

$$\begin{aligned}&\frac{1}{I}\sum_{i\in[I]} \int_\Xi Z(\mathbf{x}, \xi) \, \mathbb{P}_i(\mathrm{d}\xi) \\ &\quad = \frac{1}{I}\sum_{i\in[I]} \int_\Xi (\boldsymbol{\mathcal{T}}(\mathbf{x})\xi + \hbar(\mathbf{x}))^\top \pi(\xi) \, \mathbb{P}_i(\mathrm{d}\xi) \\ &\quad = \frac{1}{I}\sum_{i\in[I]} [\mathrm{tr}(\boldsymbol{\mathcal{T}}(\mathbf{x})\boldsymbol{\Upsilon}_i) + \hbar(\mathbf{x})^\top \gamma_i],\end{aligned}$$

where the first equality follows from the observation that $(\boldsymbol{\mathcal{T}}(\mathbf{x})\xi + \hbar(\mathbf{x}))^\top \pi(\xi) = (\mathbf{T}(\mathbf{x})\xi + \mathbf{h}(\mathbf{x}))^\top \mathbf{p}(\xi)$ for all $\xi \in \Xi$. Note that the rightmost term in the above equation corresponds to the objective value of the candidate solution (15) in (14). We have thus shown that from any feasible solution $\{\mathbb{P}_i\}_{i\in[I]}$ to the moment problem (5) we can construct a feasible solution $\{(\mu_i, \gamma_i, \boldsymbol{\Omega}_i, \boldsymbol{\Gamma}_i, \boldsymbol{\Upsilon}_i)\}_{i\in[I]}$ to the completely positive program (14) that attains the same objective value. This demonstrates that $\mathcal{Z}(\mathbf{x}) \leqslant \underline{\mathcal{Z}}(\mathbf{x})$.

To prove the converse inequality, consider any feasible solution $\{(\mu_i, \gamma_i, \boldsymbol{\Omega}_i, \boldsymbol{\Gamma}_i, \boldsymbol{\Upsilon}_i)\}_{i\in[I]}$ to (14), which gives rise to a moment matrix with the following completely positive decomposition:

$$\begin{bmatrix} \boldsymbol{\Omega}_i & \boldsymbol{\Upsilon}_i & \mu_i \\ \boldsymbol{\Upsilon}_i^\top & \boldsymbol{\Gamma}_i & \gamma_i \\ \mu_i^\top & \gamma_i^\top & 1 \end{bmatrix} = \sum_{l\in\mathscr{L}_i} \begin{bmatrix} \chi_{il} \\ \eta_{il} \\ \alpha_{il} \end{bmatrix} \begin{bmatrix} \chi_{il} \\ \eta_{il} \\ \alpha_{il} \end{bmatrix}^\top, \tag{16}$$

where $\mathscr{L}_i$ is a finite index set, while $\chi_{il} \in \mathbb{R}_+^K$, $\eta_{il} \in \mathbb{R}_+^{M+J}$, and $\alpha_{il} \in \mathbb{R}_+$ for every $l \in \mathscr{L}_i$. Partitioning $\mathscr{L}_i$ into $\mathscr{L}_i^+ = \{l \in \mathscr{L}: \alpha_{il} > 0\}$ and $\mathscr{L}_i^0 = \{l \in \mathscr{L}: \alpha_{il} = 0\}$, the decomposition (16) reduces to

$$\begin{aligned}\begin{bmatrix} \boldsymbol{\Omega}_i & \boldsymbol{\Upsilon}_i & \mu_i \\ \boldsymbol{\Upsilon}_i^\top & \boldsymbol{\Gamma}_i & \gamma_i \\ \mu_i^\top & \gamma_i^\top & 1 \end{bmatrix} &= \sum_{l\in\mathscr{L}_i^+} \begin{bmatrix} \chi_{il}\chi_{il}^\top & \chi_{il}\eta_{il}^\top & \alpha_{il}\chi_{il} \\ \eta_{il}\chi_{il}^\top & \eta_{il}\eta_{il}^\top & \alpha_{il}\eta_{il} \\ \alpha_{il}\chi_{il}^\top & \alpha_{il}\eta_{il}^\top & \alpha_{il}^2 \end{bmatrix} \\ &\quad + \sum_{l\in\mathscr{L}_i^0} \begin{bmatrix} \chi_{il}\chi_{il}^\top & \chi_{il}\eta_{il}^\top & \mathbf{0} \\ \eta_{il}\chi_{il}^\top & \eta_{il}\eta_{il}^\top & \mathbf{0} \\ \mathbf{0}^\top & \mathbf{0}^\top & 0 \end{bmatrix}. \end{aligned} \tag{17}$$

Next, we construct a sequence of discrete distributions $\mathbb{P}_i^\kappa$, $i \in [I]$, parametrized by $\kappa \in [0, 1]$, that satisfy

$$\left.\begin{aligned}\mathbb{P}_i^\kappa\left(\tilde{\xi} = \frac{\chi_{il}}{\alpha_{il}}\right) &= (1-\kappa^2)\alpha_{il}^2 && \forall l \in \mathscr{L}_i^+ \\ \mathbb{P}_i^\kappa\left(\tilde{\xi} = \hat{\xi}_i + \frac{1}{\kappa}\sqrt{|\mathscr{L}_i^0|}\chi_{il}\right) &= \frac{\kappa^2}{|\mathscr{L}_i^0|} && \forall l \in \mathscr{L}_i^0\end{aligned}\right\} \quad \forall i \in [I].$$

Observe that each $\mathbb{P}_i^\kappa$ is indeed a probability distribution since $\sum_{l\in\mathscr{L}_i^+} \alpha_{il}^2 = 1$ because of (17). Lemma 2 implies that $\chi_{il}/\alpha_{il} \in \Xi$ for every $l \in \mathscr{L}_i^+$ and $\hat{\xi}_i + (1/\kappa)\cdot\sqrt{|\mathscr{L}_i^0|}\chi_{il} \in \Xi$ for every $l \in \mathscr{L}_i^0$. Thus, $\mathbb{P}_i^\kappa$ is supported on Ξ. We further have

$$\begin{aligned}&\frac{1}{I}\sum_{i\in[I]} \mathbb{E}_{\mathbb{P}_i^\kappa}[\|\tilde{\xi} - \hat{\xi}_i\|_2^2] \\ &\quad = \frac{1}{I}\sum_{i\in[I]} \left[\sum_{l\in\mathscr{L}_i^+} (1-\kappa^2)\alpha_{il}^2 \|\chi_{il}/\alpha_{il} - \hat{\xi}_i\|_2^2 \right. \\ &\qquad \left. + \sum_{l\in\mathscr{L}_i^0} \frac{\kappa^2}{|\mathscr{L}_i^0|} \left\|\hat{\xi}_i + \frac{1}{\kappa}\sqrt{|\mathscr{L}_i^0|}\chi_{il} - \hat{\xi}_i\right\|_2^2\right] \\ &\quad \leqslant \frac{1}{I}\sum_{i\in[I]} \left[\sum_{l\in\mathscr{L}_i^+} (\chi_{il}^\top\chi_{il} - 2\hat{\xi}_i^\top(\alpha_{il}\chi_{il})) + \hat{\xi}_i^\top\hat{\xi}_i + \sum_{l\in\mathscr{L}_i^0} \chi_{il}^\top\chi_{il}\right] \\ &\quad = \frac{1}{I}\sum_{i\in[I]} [\mathrm{tr}(\boldsymbol{\Omega}_i) - 2\hat{\xi}_i^\top\mu_i + \hat{\xi}_i^\top\hat{\xi}_i] \leqslant \epsilon^2,\end{aligned}$$

where the first inequality holds since $1 - \kappa^2 \leqslant 1$, the second equality follows from the decomposition (17), and the last inequality follows from the penultimate constraint in (14). Thus, the distributions $\mathbb{P}_i^\kappa$, $i \in [I]$, are feasible in the generalized moment problem (5). We next construct feasible solutions for the dual recourse problem (4). For any $i \in [I]$ and $l \in \mathscr{L}_i$, we define ρ_{il} as the vector of the first M elements of $\eta_{il} \in \mathbb{R}_+^{M+J}$. Lemma 2 implies that $\mathbf{Q}\chi_{il}/\alpha_{il} + \mathbf{q} = \mathbf{W}^\top\rho_{il}/\alpha_{il}$ for every $l \in \mathscr{L}_i^+$. Thus, $\mathbf{p}_{il} = \rho_{il}/\alpha_{il}$ is feasible in the dual recourse problem (4) at $\xi = \chi_{il}/\alpha_{il}$ for $l \in \mathscr{L}_i^+$. Next, for any $i \in [I]$, let $\hat{\mathbf{p}}_i$ be a feasible solution to the dual recourse problem (4) at $\xi = \hat{\xi}_i$, which exists because problem (1) has sufficiently expensive recourse. Hence, we have $\mathbf{Q}\hat{\xi}_i + \mathbf{q} = \mathbf{W}^\top\hat{\mathbf{p}}_i$. Lemma 2 further implies that $\mathbf{Q}\chi_{il} = \mathbf{W}^\top\rho_{il}$ for every $l \in \mathscr{L}_i^0$. Combining the last two equalities yields

$$\mathbf{Q}\left(\hat{\xi}_i + \frac{1}{\kappa}\sqrt{|\mathscr{L}_i^0|}\chi_{il}\right) + \mathbf{q} = \mathbf{W}^\top\left(\hat{\mathbf{p}}_i + \frac{1}{\kappa}\sqrt{|\mathscr{L}_i^0|}\rho_{il}\right) \quad \forall l \in \mathscr{L}_i^0.$$

Thus, $\mathbf{p}_{il} = \hat{\mathbf{p}}_i + (1/\kappa)\sqrt{|\mathscr{L}_i^0|}\boldsymbol{\rho}_{il}$ constitutes a feasible solution in the dual recourse problem (4) at $\boldsymbol{\xi} = \hat{\boldsymbol{\xi}}_i + (1/\kappa)\sqrt{|\mathscr{L}_i^0|}\boldsymbol{\chi}_{il}$ for $l \in \mathscr{L}_i^0$. In summary, we have

$$Z\left(\mathbf{x}, \frac{\boldsymbol{\chi}_{il}}{\alpha_{il}}\right) = Z_{\mathrm{d}}\left(\mathbf{x}, \frac{\boldsymbol{\chi}_{il}}{\alpha_{il}}\right) \geqslant \left(\mathbf{T}(\mathbf{x})\frac{\boldsymbol{\chi}_{il}}{\alpha_{il}} + \mathbf{h}(\mathbf{x})\right)^\top \mathbf{p}_{il} \quad \forall l \in \mathscr{L}_i^+$$

and

$$\begin{aligned} &Z\left(\mathbf{x}, \hat{\boldsymbol{\xi}}_i + \frac{1}{\kappa}\sqrt{|\mathscr{L}_i^0|}\boldsymbol{\chi}_{il}\right) \\ &\quad = Z_{\mathrm{d}}\left(\mathbf{x}, \hat{\boldsymbol{\xi}}_i + \frac{1}{\kappa}\sqrt{|\mathscr{L}_i^0|}\boldsymbol{\chi}_{il}\right) \\ &\quad \geqslant \left(\mathbf{T}(\mathbf{x})\left(\hat{\boldsymbol{\xi}}_i + \frac{1}{\kappa}\sqrt{|\mathscr{L}_i^0|}\boldsymbol{\chi}_{il}\right) + \mathbf{h}(\mathbf{x})\right)^\top \mathbf{p}_{il} \quad \forall l \in \mathscr{L}_i^0. \end{aligned}$$

Using these estimates, we can now bound the objective value of the discrete distributions $\mathbb{P}_i^\kappa$, $i \in [I]$, in (5). Specifically, we obtain

$$\begin{aligned} &\frac{1}{I}\sum_{i\in[I]} \mathbb{E}_{\mathbb{P}_i^\kappa}[Z(\mathbf{x}, \tilde{\boldsymbol{\xi}})] \\ &= \frac{1}{I}\sum_{i\in[I]}\left[\sum_{l\in\mathscr{L}_i^+}(1-\kappa^2)\alpha_{il}^2 Z\left(\mathbf{x}, \frac{\boldsymbol{\chi}_{il}}{\alpha_{il}}\right)\right. \\ &\qquad \left. + \sum_{l\in\mathscr{L}_i^0}\frac{\kappa^2}{|\mathscr{L}_i^0|} Z\left(\mathbf{x}, \hat{\boldsymbol{\xi}}_i + \frac{\sqrt{|\mathscr{L}_i^0|}}{\kappa}\boldsymbol{\chi}_{il}\right)\right] \\ &\geqslant \frac{1}{I}\sum_{i\in[I]}\left[\sum_{l\in\mathscr{L}_i^+}(1-\kappa^2)\alpha_{il}^2\left(\mathbf{T}(\mathbf{x})\frac{\boldsymbol{\chi}_{il}}{\alpha_{il}} + \mathbf{h}(\mathbf{x})\right)^\top \mathbf{p}_{il}\right. \\ &\qquad \left. + \sum_{l\in\mathscr{L}_i^0}\frac{\kappa^2}{|\mathscr{L}_i^0|}\left(\mathbf{T}(\mathbf{x})\left(\hat{\boldsymbol{\xi}}_i + \frac{\sqrt{|\mathscr{L}_i^0|}}{\kappa}\boldsymbol{\chi}_{il}\right) + \mathbf{h}(\mathbf{x})\right)^\top \mathbf{p}_{il}\right] \\ &= \frac{1}{I}\sum_{i\in[I]}\left[\sum_{l\in\mathscr{L}_i^+}(1-\kappa^2)\alpha_{il}^2\left(\boldsymbol{\mathcal{T}}(\mathbf{x})\frac{\boldsymbol{\chi}_{il}}{\alpha_{il}} + \hbar(\mathbf{x})\right)^\top \frac{\boldsymbol{\eta}_{il}}{\alpha_{il}}\right. \\ &\qquad + \sum_{l\in\mathscr{L}_i^0}\frac{\kappa^2}{|\mathscr{L}_i^0|}\left(\boldsymbol{\mathcal{T}}(\mathbf{x})\left(\hat{\boldsymbol{\xi}}_i + \frac{\sqrt{|\mathscr{L}_i^0|}}{\kappa}\boldsymbol{\chi}_{il}\right) + \hbar(\mathbf{x})\right)^\top \\ &\qquad \left.\cdot\left([\hat{\mathbf{p}}_i^\top\ \mathbf{0}]^\top + \frac{\sqrt{|\mathscr{L}_i^0|}}{\kappa}\boldsymbol{\eta}_{il}\right)\right] \\ &= \frac{1}{I}\sum_{i\in[I]}\left[\sum_{l\in\mathscr{L}_i^+}(1-\kappa^2)[\mathrm{tr}(\boldsymbol{\mathcal{T}}(\mathbf{x})\boldsymbol{\chi}_{il}\boldsymbol{\eta}_{il}^\top) + \hbar(\mathbf{x})^\top(\alpha_{il}\boldsymbol{\eta}_{il})]\right. \\ &\qquad \left. + \sum_{l\in\mathscr{L}_i^0}\mathrm{tr}(\boldsymbol{\mathcal{T}}(\mathbf{x})\boldsymbol{\chi}_{il}\boldsymbol{\eta}_{il}^\top)\right] \\ &\quad + \frac{1}{I}\sum_{i\in[I]}\left[\sum_{l\in\mathscr{L}_i^0}\frac{\kappa^2}{|\mathscr{L}_i^0|}\left(\left(\mathbf{T}(\mathbf{x})\left(\hat{\boldsymbol{\xi}}_i + \frac{\sqrt{|\mathscr{L}_i^0|}}{\kappa}\boldsymbol{\chi}_{il}\right) + \mathbf{h}(\mathbf{x})\right)^\top \hat{\mathbf{p}}_i\right.\right. \\ &\qquad \left.\left. + (\boldsymbol{\mathcal{T}}(\mathbf{x})\hat{\boldsymbol{\xi}}_i + \hbar(\mathbf{x}))^\top \frac{\sqrt{|\mathscr{L}_i^0|}}{\kappa}\boldsymbol{\eta}_{il}\right)\right]. \end{aligned}$$

Together with the decomposition (17), the above estimate implies that

$$\lim_{\kappa\downarrow 0}\frac{1}{I}\sum_{i\in[I]}\mathbb{E}_{\mathbb{P}_i^\kappa}[Z(\mathbf{x}, \tilde{\boldsymbol{\xi}})] \geqslant \frac{1}{I}\sum_{i\in[I]}\mathrm{tr}(\boldsymbol{\mathcal{T}}(\mathbf{x})\boldsymbol{\Upsilon}_i) + \hbar(\mathbf{x})^\top\boldsymbol{\gamma}_i.$$

We have therefore shown that any feasible solution to (14) can be used to construct a sequence of feasible solutions to the generalized moment problem (5) that asymptotically attain a (weakly) larger objective value. This demonstrates that $\mathscr{Z}(\mathbf{x}) \geqslant \underline{\mathscr{Z}}(\mathbf{x})$. Thus the claim follows. □

The proof of Theorem 3 relies on the following lemma, which is inspired by lemma 2.2 in Burer (2009) and proposition 3.1 in Natarajan et al. (2011).

Lemma 2. *If* $\{(\boldsymbol{\mu}_i, \boldsymbol{\gamma}_i, \boldsymbol{\Omega}_i, \boldsymbol{\Gamma}_i, \boldsymbol{\Upsilon}_i)\}_{i\in[I]}$ *is feasible in* (14) *and if* $(\boldsymbol{\chi}_{il}, \boldsymbol{\eta}_{il}, \alpha_{il}) \in \mathbb{R}_+^K \times \mathbb{R}_+^{M+J} \times \mathbb{R}_+$, $l \in \mathscr{L}_i$, *satisfies* (16) *for some* $i \in [I]$, *then*

$$\begin{aligned} \boldsymbol{\chi}_{il}/\alpha_{il} \in \Xi \quad &\textit{and} \\ \mathbf{Q}(\boldsymbol{\chi}_{il}/\alpha_{il}) + \mathbf{q} = \mathbf{W}^\top(\boldsymbol{\rho}_{il}/\alpha_{il}) \quad &\forall l \in \mathscr{L}_i^+, \end{aligned}$$

while

$$\boldsymbol{\chi}_{il} \in \mathrm{recc}(\Xi) \quad \textit{and} \quad \mathbf{Q}\boldsymbol{\chi}_{il} = \mathbf{W}^\top\boldsymbol{\rho}_{il} \quad \forall l \in \mathscr{L}_i^0,$$

where $\mathrm{recc}(\Xi) = \{\boldsymbol{\xi} \in \mathbb{R}_+^K : \mathbf{S}\boldsymbol{\xi} \leqslant \mathbf{0}\}$ *is the recession cone of* Ξ *and* $\boldsymbol{\rho}_{il}$ *is the vector of the first* M *elements of* $\boldsymbol{\eta}_{il}$.

Proof. We substitute the decomposition (16) into the constraints of problem (14) to obtain

$$\sum_{l\in\mathscr{L}_i}\alpha_{il}(\boldsymbol{\mathcal{W}}_{:j}^\top\boldsymbol{\eta}_{il} - \boldsymbol{\mathcal{Q}}_{j:}^\top\boldsymbol{\chi}_{il}) = q_j \tag{18}$$

and

$$\sum_{l\in\mathscr{L}_i}(\boldsymbol{\mathcal{W}}_{:j}^\top\boldsymbol{\eta}_{il} - \boldsymbol{\mathcal{Q}}_{j:}^\top\boldsymbol{\chi}_{il})^2 = q_j^2 \tag{19}$$

for every $j \in [N_2 + J]$. Squaring (18) and eliminating q_j^2 by using (19) yields

$$\begin{aligned} &\left(\sum_{l\in\mathscr{L}_i}\alpha_{il}\left(\boldsymbol{\mathcal{W}}_{:j}^\top\boldsymbol{\eta}_{il} - \boldsymbol{\mathcal{Q}}_{j:}^\top\boldsymbol{\chi}_{il}\right)\right)^2 \\ &\quad = \sum_{l\in\mathscr{L}_i}\left(\boldsymbol{\mathcal{W}}_{:j}^\top\boldsymbol{\eta}_{il} - \boldsymbol{\mathcal{Q}}_{j:}^\top\boldsymbol{\chi}_{il}\right)^2 \\ &\quad = \left(\sum_{l\in\mathscr{L}_i}\alpha_{il}^2\right)\sum_{l\in\mathscr{L}_i}\left(\boldsymbol{\mathcal{W}}_{:j}^\top\boldsymbol{\eta}_{il} - \boldsymbol{\mathcal{Q}}_{j:}^\top\boldsymbol{\chi}_{il}\right)^2. \end{aligned}$$

Here, the second equality follows from the fact that $\sum_{l\in\mathscr{L}}\alpha_{il}^2 = 1$. The tightness condition of the Cauchy–Schwarz inequality therefore implies that there exists $\tau \in \mathbb{R}$ with

$$\boldsymbol{\mathcal{W}}_{:j}^\top\boldsymbol{\eta}_{il} - \boldsymbol{\mathcal{Q}}_{j:}^\top\boldsymbol{\chi}_{il} = \tau\alpha_{il} \quad \forall l \in \mathscr{L}_i. \tag{20}$$

Thus, we have

$$\boldsymbol{\mathcal{W}}_{:j}^{\top}\boldsymbol{\eta}_{il} - \boldsymbol{\mathcal{Q}}_{j:}^{\top}\boldsymbol{\chi}_{il} = 0 \quad \forall\, l \in \mathscr{L}_i^0,$$

which confirms the second claim. Next, we observe from (18) and (20) that

$$q_j = \sum_{l\in\mathscr{L}_i} \alpha_{il}(\boldsymbol{\mathcal{W}}_{:j}^{\top}\boldsymbol{\eta}_{il} - \boldsymbol{\mathcal{Q}}_{j:}^{\top}\boldsymbol{\chi}_{il}) = \sum_{l\in\mathscr{L}_i} \tau\alpha_{il}^2 = \tau. \tag{21}$$

Replacing τ with q_j in (20) and using the fact that $\alpha_{il} > 0$ for $l \in \mathscr{L}_i^+$ yields

$$\boldsymbol{\mathcal{W}}_{:j}^{\top}(\boldsymbol{\eta}_{il}/\alpha_{il}) - \boldsymbol{\mathcal{Q}}_{j:}^{\top}(\boldsymbol{\chi}_{il}/\alpha_{il}) = q_j \quad \forall\, l \in \mathscr{L}_i^+,$$

which establishes the first claim. □

3.3. A Copositive Reformulation of Problem (1)

So far, we have seen that $\mathscr{Z}(\mathbf{x}) = \underline{\mathscr{Z}}(\mathbf{x}) \leqslant \bar{\mathscr{Z}}(\mathbf{x})$. Unfortunately, as we exemplify below, the duality gap between $\underline{\mathscr{Z}}(\mathbf{x})$ and $\bar{\mathscr{Z}}(\mathbf{x})$ can be strictly positive.

Example 1 (Infinite Duality Gap)**.** Consider the following pair of primal and dual recourse problems:

$$Z(x,\xi) = \inf_{y\in\mathbb{R}}\{(\xi-1)y\colon \xi - 1 \leqslant 0y\} \quad \text{and}$$
$$Z_d(x,\xi) = \sup_{p\in\mathbb{R}_+}\{(\xi-1)p\colon \xi - 1 = 0p\},$$

respectively, and set $\Xi = \{\xi \in \mathbb{R}_+\colon \xi \leqslant 1, -\xi \leqslant -1\} = \{1\}$. Assume that there is only one sample $\hat{\xi}_1 = 1$ and that the Wasserstein radius is set to $\epsilon = 1$. Note that both linear programs are feasible with the same optimal value 0 for $\xi = 1$. Theorem 3 therefore implies that $\underline{\mathscr{Z}}(\mathbf{x}) = \mathscr{Z}(\mathbf{x}) = 0$.

Under the current setting, the extended recourse parameters defined in (11) are given by

$$\boldsymbol{\mathcal{Q}} = \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix}, \quad q = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}, \quad \boldsymbol{\mathcal{T}}(\mathbf{x}) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$
$$\boldsymbol{\mathcal{W}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad \text{and} \quad h(\mathbf{x}) = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}.$$

Hence, the copositive program (10) simplifies to

$$\begin{aligned} \bar{\mathscr{Z}}(\mathbf{x}) = \inf\ & \{s - \psi_1 - \psi_2 + \psi_3 + \phi_1 + \phi_2 + \phi_3\} \\ \text{s.t.}\ & \lambda \in \mathbb{R}_+, s \in \mathbb{R}, \boldsymbol{\psi}, \boldsymbol{\phi} \in \mathbb{R}^3 \\ & \begin{bmatrix} \lambda + \phi_1 + \phi_2 + \phi_3 & -\frac{1}{2} & \phi_2 \\ -\frac{1}{2} & 0 & 0 \\ \phi_2 & 0 & \phi_2 \\ -\phi_3 & 0 & 0 \\ -\lambda - \frac{1}{2}(\psi_1 + \psi_2 - \psi_3) & \frac{1}{2} & -\frac{1}{2}\psi_2 \end{bmatrix} \\ & \qquad \begin{bmatrix} -\phi_3 & -\lambda - \frac{1}{2}(\psi_1 + \psi_2 - \psi_3) \\ 0 & \frac{1}{2} \\ 0 & -\frac{1}{2}\psi_2 \\ \phi_3 & -\frac{1}{2}\psi_3 \\ -\frac{1}{2}\psi_3 & s \end{bmatrix} \succeq_{\mathscr{C}} \mathbf{0}. \end{aligned} \tag{22}$$

However, multiplying the copositive constraint from both sides with the vector $[\theta\ 1\ 0\ 0\ 0]^{\top}$, $\theta \geqslant 0$, implies that $(\lambda + \phi_1 + \phi_2 + \phi_3)\theta^2 - \theta \geqslant 0$ for all $\theta \geqslant 0$. As no values of λ, ϕ_1, ϕ_2, or ϕ_3 can satisfy this inequality, problem (22) is infeasible (i.e., $\bar{\mathscr{Z}}(\mathbf{x}) = +\infty$). Thus, there is an infinite duality gap between $\underline{\mathscr{Z}}(\mathbf{x})$ and $\bar{\mathscr{Z}}(\mathbf{x})$.

Even though $\underline{\mathscr{Z}}(\mathbf{x})$ and $\bar{\mathscr{Z}}(\mathbf{x})$ may differ, one can prove that $\underline{\mathscr{Z}}(\mathbf{x}) = \mathscr{Z}(\mathbf{x}) = \bar{\mathscr{Z}}(\mathbf{x})$ if the two-stage distributionally robust linear program (1) has complete recourse. To show this, we first prove two lemmas.

Lemma 3. *If problem* (1) *has complete recourse, then* $\mathbf{W}\mathbf{W}^{\top} \succ_{\mathscr{C}} \mathbf{0}$.

Proof. The complete recourse property is equivalent to the unboundedness of the linear program

$$\begin{aligned} \text{maximize}\ & z \\ \text{subject to}\ & \mathbf{y} \in \mathbb{R}^{N_2},\ z \in \mathbb{R}, \\ & \mathbf{W}\mathbf{y} \geqslant z\mathbf{e}, \end{aligned}$$

whose dual linear program is given by

$$\begin{aligned} \text{minimize}\ & 0 \\ \text{subject to}\ & \boldsymbol{\lambda} \in \mathbb{R}_+^M, \\ & \mathbf{e}^{\top}\boldsymbol{\lambda} = 1,\ \mathbf{W}^{\top}\boldsymbol{\lambda} = \mathbf{0}. \end{aligned}$$

As the primal problem is unbounded, the dual problem is infeasible by weak duality, implying that $\mathbf{W}^{\top}\boldsymbol{\lambda} \neq \mathbf{0}$ for all $\boldsymbol{\lambda} \in \mathbb{R}_+^M$ such that $\mathbf{e}^{\top}\boldsymbol{\lambda} = 1$. By rescaling, we thus have $\boldsymbol{\lambda}^{\top}\mathbf{W}\mathbf{W}^{\top}\boldsymbol{\lambda} > 0$ for all $\boldsymbol{\lambda} \in \mathbb{R}_+^M$ such that $\boldsymbol{\lambda} \neq \mathbf{0}$. This implies that $\mathbf{W}\mathbf{W}^{\top}$ lies in the interior of the copositive cone $\mathscr{C}$. □

Lemma 4 (Copositive Schur Complements)**.** *Consider the symmetric matrix*

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^{\top} & \mathbf{C} \end{bmatrix},$$

with $\mathbf{A} \succ \mathbf{0}$. *We then have* $\mathbf{M} \succ_{\mathscr{C}} \mathbf{0}$ *if* $\mathbf{C} - \mathbf{B}^{\top}\mathbf{A}^{-1}\mathbf{B} \succ_{\mathscr{C}} \mathbf{0}$.

Proof. Multiplying the matrix $\mathbf{M}$ from both sides with a nonnegative vector $[\boldsymbol{\xi}^{\top}\ \boldsymbol{\rho}^{\top}]^{\top} \in \mathbb{R}_+^{K+M}$ satisfying $\mathbf{e}^{\top}\boldsymbol{\xi} + \mathbf{e}^{\top}\boldsymbol{\rho} = 1$, we obtain

$$\begin{aligned} [\boldsymbol{\xi}^{\top}\ \boldsymbol{\rho}^{\top}]\mathbf{M}[\boldsymbol{\xi}^{\top}\ \boldsymbol{\rho}^{\top}]^{\top} &= \boldsymbol{\xi}^{\top}\mathbf{A}\boldsymbol{\xi} + 2\boldsymbol{\xi}^{\top}\mathbf{B}\boldsymbol{\rho} + \boldsymbol{\rho}^{\top}\mathbf{C}\boldsymbol{\rho} \\ &= (\boldsymbol{\xi} + \mathbf{A}^{-1}\mathbf{B}\boldsymbol{\rho})^{\top}\mathbf{A}(\boldsymbol{\xi} + \mathbf{A}^{-1}\mathbf{B}\boldsymbol{\rho}) \\ &\quad + \boldsymbol{\rho}^{\top}(\mathbf{C} - \mathbf{B}^{\top}\mathbf{A}^{-1}\mathbf{B})\boldsymbol{\rho}. \end{aligned}$$

Since $\mathbf{A} \succ \mathbf{0}$, the term $(\boldsymbol{\xi} + \mathbf{A}^{-1}\mathbf{B}\boldsymbol{\rho})^{\top}\mathbf{A}(\boldsymbol{\xi} + \mathbf{A}^{-1}\mathbf{B}\boldsymbol{\rho})$ is nonnegative. If $\boldsymbol{\rho} = \mathbf{0}$, then we have $\mathbf{e}^{\top}\boldsymbol{\xi} = 1$, which implies that this term is positive. If $\boldsymbol{\rho} \neq \mathbf{0}$, then the assumption $\mathbf{C} - \mathbf{B}^{\top}\mathbf{A}^{-1}\mathbf{B} \succ_{\mathscr{C}} \mathbf{0}$ implies that the term $\boldsymbol{\rho}^{\top}(\mathbf{C} - \mathbf{B}^{\top}\mathbf{A}^{-1}\mathbf{B})\boldsymbol{\rho}$ is positive. In both cases, by rescaling, we find that $[\boldsymbol{\xi}^{\top}\ \boldsymbol{\rho}^{\top}]\mathbf{M}[\boldsymbol{\xi}^{\top}\ \boldsymbol{\rho}^{\top}]^{\top} > 0$ for all $\boldsymbol{\xi} \in \mathbb{R}_+^K$ and all $\boldsymbol{\rho} \in \mathbb{R}_+^M$ such that $[\boldsymbol{\xi}^{\top}\ \boldsymbol{\rho}^{\top}]^{\top} \neq \mathbf{0}$. Hence, $\mathbf{M} \succ_{\mathscr{C}} \mathbf{0}$. □

Lemmas 3 and 4 enable us to prove the following exactness result.

Theorem 4. *If problem (1) has complete recourse, then* $\mathcal{Z}(\mathbf{x}) = \underline{\mathcal{Z}}(\mathbf{x}) = \bar{\mathcal{Z}}(\mathbf{x})$ *for any fixed* $\mathbf{x} \in \mathcal{X}$.

Proof. We already know from Theorem 3 that $\mathcal{Z}(\mathbf{x}) = \underline{\mathcal{Z}}(\mathbf{x})$. To show that $\bar{\mathcal{Z}}(\mathbf{x}) = \underline{\mathcal{Z}}(\mathbf{x})$, it suffices to prove strong duality between problems (10) and (14). Specifically, we will construct a Slater point $(\lambda^s, (s_i^s, \boldsymbol{\phi}_i^s, \boldsymbol{\psi}_i^s)_{i\in[I]})$ for problem (10). To this end, we first set $\boldsymbol{\phi}_i^s = \mathbf{e}$ and $\boldsymbol{\psi}_i^s = \mathbf{0}$ for all $i \in [I]$. Using this solution, the ith constraint matrix in (10) can be decomposed as

$$\begin{bmatrix} \lambda^s \mathbb{I} & -\frac{1}{2}\mathbf{T}(\mathbf{x})^\top - \mathbf{Q}^\top \mathbf{W}^\top & \mathbf{S}^\top & \mathbf{0} \\ -\frac{1}{2}\mathbf{T}(\mathbf{x}) - \mathbf{W}\mathbf{Q} & \mathbf{W}\mathbf{W}^\top & \mathbf{0} & -\frac{1}{2}\mathbf{h}(\mathbf{x}) \\ \mathbf{S} & \mathbf{0} & \mathbb{I} & \mathbf{0} \\ \mathbf{0}^\top & -\frac{1}{2}\mathbf{h}(\mathbf{x})^\top & \mathbf{0}^\top & \lambda^s \end{bmatrix} + \begin{bmatrix} \boldsymbol{\mathcal{Q}}^\top \boldsymbol{\mathcal{Q}} & \mathbf{0} & \mathbf{0} & -\lambda^s \hat{\boldsymbol{\xi}}_i \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\lambda^s \hat{\boldsymbol{\xi}}_i^\top & \mathbf{0}^\top & \mathbf{0}^\top & s_i^s - \lambda^s \end{bmatrix}. \tag{23}$$

Next, we select s_i^s large enough to ensure that the right matrix in (23) is positive semidefinite. In this case, a Slater point can be obtained by ensuring that the left matrix is strictly copositive. As problem (1) has complete recourse, Lemma 3 is applicable and implies that

$$\begin{bmatrix} \mathbf{W}\mathbf{W}^\top & \mathbf{0} \\ \mathbf{0} & \mathbb{I} \end{bmatrix} \succ_{\mathscr{C}} \mathbf{0}.$$

Moreover, Lemma 4 implies that the left matrix in (23) is strictly copositive if

$$\begin{bmatrix} \mathbf{W}\mathbf{W}^\top & \mathbf{0} \\ \mathbf{0} & \mathbb{I} \end{bmatrix} \succ_{\mathscr{C}} \frac{1}{\lambda^s} \begin{bmatrix} -\frac{1}{2}\mathbf{T}(\mathbf{x})^\top + \mathbf{Q}^\top \mathbf{W}^\top & \mathbf{S}^\top \\ -\frac{1}{2}\mathbf{h}(\mathbf{x})^\top & \mathbf{0}^\top \end{bmatrix}^\top \cdot \begin{bmatrix} -\frac{1}{2}\mathbf{T}(\mathbf{x})^\top + \mathbf{Q}^\top \mathbf{W}^\top & \mathbf{S}^\top \\ -\frac{1}{2}\mathbf{h}(\mathbf{x})^\top & \mathbf{0}^\top \end{bmatrix},$$

which is true whenever λ^s is sufficiently large. We have therefore constructed a strictly feasible solution to (10). Thus, $\bar{\mathcal{Z}}(\mathbf{x}) = \underline{\mathcal{Z}}(\mathbf{x}) = Z(\mathbf{x})$ by strong conic duality. □

Theorem 4 implies that if problem (1) has complete recourse, then it is equivalent to the copositive minimization problem obtained by replacing $\mathcal{Z}(\mathbf{x})$ in (1) with $\bar{\mathcal{Z}}(\mathbf{x})$. Conversely, if problem (1) fails to have complete recourse, it may only be possible to approximate $\mathcal{Z}(\mathbf{x})$ by the optimal value of a copositive minimization problem. To show this, we construct a relaxation of problem (10) parameterized by $\delta \geqslant 0$:

$$\bar{\mathcal{Z}}_\delta(\mathbf{x}) = \inf \left\{ \epsilon^2 \lambda + \frac{1}{I} \sum_{i\in[I]} \left[s_i + \mathbf{q}^\top \boldsymbol{\psi}_i - \lambda \|\hat{\boldsymbol{\xi}}_i\|_2^2 + \sum_{j\in[N_2+J]} \phi_{ij} q_j^2 \right] \right\}$$
$$\text{s.t. } \lambda \in \mathbb{R}_+, s_i \in \mathbb{R}, \boldsymbol{\psi}_i, \boldsymbol{\phi}_i \in \mathbb{R}^{N_2+J} \quad \forall i \in [I],$$
$$\begin{bmatrix} \lambda \mathbb{I} + \boldsymbol{\mathcal{Q}}^\top \operatorname{diag}(\boldsymbol{\phi}_i) \boldsymbol{\mathcal{Q}} & -\frac{1}{2}\boldsymbol{\mathcal{T}}(\mathbf{x})^\top - \boldsymbol{\mathcal{Q}}^\top \operatorname{diag}(\boldsymbol{\phi}_i) \boldsymbol{\mathcal{W}}^\top & -\lambda \hat{\boldsymbol{\xi}}_i - \frac{1}{2}\boldsymbol{\mathcal{Q}}^\top \boldsymbol{\psi}_i \\ -\frac{1}{2}\boldsymbol{\mathcal{T}}(\mathbf{x}) - \boldsymbol{\mathcal{W}} \operatorname{diag}(\boldsymbol{\phi}_i) \boldsymbol{\mathcal{Q}} & \boldsymbol{\mathcal{W}} \operatorname{diag}(\boldsymbol{\phi}_i) \boldsymbol{\mathcal{W}}^\top + \delta \mathbb{I} & \frac{1}{2}(\boldsymbol{\mathcal{W}} \boldsymbol{\psi}_i - \boldsymbol{\hbar}(\mathbf{x})) \\ (-\lambda \hat{\boldsymbol{\xi}}_i - \frac{1}{2}\mathbf{Q}^\top \boldsymbol{\psi}_i)^\top & \frac{1}{2}(\boldsymbol{\mathcal{W}} \boldsymbol{\psi}_i - \boldsymbol{\hbar}(\mathbf{x}))^\top & s_i \end{bmatrix} \succeq_{\mathscr{C}} \mathbf{0} \quad \forall i \in [I]. \tag{24}$$

Note that δ only affects the middle block of the copositive matrix. One can further show that the completely positive program dual to (24) constitutes a restriction of problem (14) with a perturbed objective function:

$$\begin{aligned} \underline{\mathcal{Z}}_\delta(\mathbf{x}) = \sup \; & \frac{1}{I} \sum_{i\in[I]} [\operatorname{tr}(\boldsymbol{\mathcal{T}}(\mathbf{x}) \boldsymbol{\Upsilon}_i) + \boldsymbol{\hbar}(\mathbf{x})^\top \boldsymbol{\gamma}_i - \delta \operatorname{tr}(\boldsymbol{\Gamma}_i)] \\ \text{s.t. } & \boldsymbol{\gamma}_i \in \mathbb{R}_+^{M+J}, \boldsymbol{\mu}_i \in \mathbb{R}_+^K, \boldsymbol{\Gamma}_i \in \mathbb{S}_+^{M+J}, \\ & \boldsymbol{\Omega}_i \in \mathbb{S}_+^K, \boldsymbol{\Upsilon}_i \in \mathbb{R}^{K \times (M+J)} \quad \forall i \in [I], \\ & \boldsymbol{\mathcal{Q}} \boldsymbol{\mu}_i + q = \boldsymbol{\mathcal{W}}^\top \boldsymbol{\gamma}_i \quad \forall i \in [I], \\ & \boldsymbol{\mathcal{Q}}_{j:}^\top \boldsymbol{\Omega}_i \boldsymbol{\mathcal{Q}}_{j:} - 2\boldsymbol{\mathcal{Q}}_{j:}^\top \boldsymbol{\Upsilon}_i \boldsymbol{\mathcal{W}}_{:j} + \boldsymbol{\mathcal{W}}_{:j}^\top \boldsymbol{\Gamma}_i \boldsymbol{\mathcal{W}}_{:j} = q_j^2 \\ & \qquad \forall i \in [I] \forall j \in [N_2 + J], \\ & \frac{1}{I} \sum_{i\in[I]} \operatorname{tr}(\boldsymbol{\Omega}_i) - 2\hat{\boldsymbol{\xi}}_i^\top \boldsymbol{\mu}_i + \hat{\boldsymbol{\xi}}_i^\top \hat{\boldsymbol{\xi}}_i \leqslant \epsilon^2, \\ & \begin{bmatrix} \boldsymbol{\Omega}_i & \boldsymbol{\Upsilon}_i & \boldsymbol{\mu}_i \\ \boldsymbol{\Upsilon}_i^\top & \boldsymbol{\Gamma}_i & \boldsymbol{\gamma}_i \\ \boldsymbol{\mu}_i^\top & \boldsymbol{\gamma}_i^\top & 1 \end{bmatrix} \succeq_{\mathscr{C}^*} \mathbf{0} \quad \forall i \in [I]. \end{aligned} \tag{25}$$

Observe that δ only affects the objective function of this dual problem.

Proposition 2. *For any fixed* $\mathbf{x} \in \mathcal{X}$, $\bar{\mathcal{Z}}_\delta(\mathbf{x}) = \underline{\mathcal{Z}}_\delta(\mathbf{x})$ *is finite for all* $\delta > 0$, *and* $\lim_{\delta \downarrow 0} \bar{\mathcal{Z}}_\delta(\mathbf{x}) = \mathcal{Z}(\mathbf{x})$.

Proof. We first show that $\bar{\mathcal{Z}}_\delta(\mathbf{x}) = \underline{\mathcal{Z}}_\delta(\mathbf{x})$ by proving strong duality between problems (24) and (25). To this end, we first construct a Slater point $(\lambda^s, (s_i^s, \boldsymbol{\phi}_i^s, \boldsymbol{\psi}_i^s)_{i\in[I]})$ for problem (24). Specifically, we set $\boldsymbol{\psi}_i^s = \mathbf{0}$ and $\boldsymbol{\phi}_i^s = \mathbf{0}$ for all $i \in [I]$, and we select λ^s satisfying $\lambda^s \mathbb{I} > (1/(4\delta)) \boldsymbol{\mathcal{T}}(\mathbf{x}) \boldsymbol{\mathcal{T}}(\mathbf{x})^\top$. This is possible because $\delta > 0$. A standard Schur complement argument then implies that for all sufficiently large $s_i^s > 0$, $i \in [I]$, we have

$$\begin{bmatrix} \lambda^s \mathbb{I} & -\frac{1}{2}\boldsymbol{\mathcal{T}}(\mathbf{x})^\top \\ -\frac{1}{2}\boldsymbol{\mathcal{T}}(\mathbf{x}) & \delta \mathbb{I} \end{bmatrix} \succ \mathbf{0}$$
$$\Rightarrow \begin{bmatrix} \lambda^s \mathbb{I} & -\frac{1}{2}\boldsymbol{\mathcal{T}}(\mathbf{x})^\top \\ -\frac{1}{2}\boldsymbol{\mathcal{T}}(\mathbf{x}) & \delta \mathbb{I} \end{bmatrix} \succ \frac{1}{s_i^s} \begin{bmatrix} \lambda^s \hat{\boldsymbol{\xi}}_i \\ \frac{1}{2}\boldsymbol{\hbar}(\mathbf{x}) \end{bmatrix} \begin{bmatrix} \lambda^s \hat{\boldsymbol{\xi}}_i \\ \frac{1}{2}\boldsymbol{\hbar}(\mathbf{x}) \end{bmatrix}^\top \quad \forall i \in [I].$$

A second Schur complement argument then ensures that

$$\begin{bmatrix} \lambda^s \mathbb{I} & -\frac{1}{2}\boldsymbol{\mathcal{T}}(\mathbf{x})^\top & -\lambda^s \hat{\boldsymbol{\xi}}_i \\ -\frac{1}{2}\boldsymbol{\mathcal{T}}(\mathbf{x}) & \delta \mathbb{I} & -\frac{1}{2}\boldsymbol{\hbar}(\mathbf{x}) \\ -\lambda^s \hat{\boldsymbol{\xi}}_i^\top & -\frac{1}{2}\boldsymbol{\hbar}(\mathbf{x})^\top & s_i^s \end{bmatrix} \succ \mathbf{0} \quad \forall i \in [I]. \tag{26}$$

Consequently, the matrix on the left-hand side of (26) is in the interior of the copositive cone $\mathscr{C}$. This proves

that $(\lambda^s, (s_i^s, \phi_i^s, \psi_i^s)_{i\in[I]})$ is indeed a Slater point for (24). Hence, $\bar{\mathcal{Z}}_\delta(\mathbf{x}) = \underline{\mathcal{Z}}_\delta(\mathbf{x})$ by strong conic duality. As problem (24) is feasible, we have $\bar{\mathcal{Z}}_\delta(\mathbf{x}) < +\infty$ for any fixed $\delta > 0$. Moreover, as problem (1) has sufficiently expensive recourse, we have $Z(\mathbf{x}, \boldsymbol{\xi}) > -\infty$ for any fixed $\mathbf{x} \in \mathscr{X}$ and $\boldsymbol{\xi} \in \Xi$. Since $\hat{\mathbb{P}}$ is nonempty, evaluating the worst-case expectation in (2) yields $\underline{\mathcal{Z}}(\mathbf{x}) = \mathcal{Z}(\mathbf{x}) > -\infty$, where the equality follows from Theorem 3. Thus, the completely positive program (14) and its restriction (25) are both feasible, implying that $\bar{\mathcal{Z}}_\delta(\mathbf{x}) = \underline{\mathcal{Z}}_\delta(\mathbf{x}) > -\infty$. This proves finiteness of $\bar{\mathcal{Z}}_\delta(\mathbf{x})$.

To prove the second claim, we observe that $\underline{\mathcal{Z}}_\delta(\mathbf{x})$ constitutes a pointwise supremum of a family of affine functions in δ. Thus, $\underline{\mathcal{Z}}_\delta(\mathbf{x})$ is convex and lower semicontinuous in δ for every fixed $\mathbf{x} \in \mathscr{X}$. Since $\underline{\mathcal{Z}}_\delta(\mathbf{x})$ is also nonincreasing in δ by construction, it is indeed right continuous. Thus, we have

$$\lim_{\delta\downarrow 0} \bar{\mathcal{Z}}_\delta(\mathbf{x}) = \lim_{\delta\downarrow 0} \underline{\mathcal{Z}}_\delta(\mathbf{x}) = \underline{\mathcal{Z}}_0(\mathbf{x}) = \mathcal{Z}(\mathbf{x}). \tag{27}$$

Here, the first equality holds because $\bar{\mathcal{Z}}_\delta(\mathbf{x}) = \underline{\mathcal{Z}}_\delta(\mathbf{x})$ for $\delta > 0$, while the second equality follows from the right continuity of $\underline{\mathcal{Z}}_\delta(\mathbf{x})$. The last equality is due to Theorem 3. This completes the proof.

The findings of this section culminate in the following main theorem.

Theorem 5. *Consider the following family of copositive programs parametrized in δ:*

minimize

$$\mathbf{c}^\top \mathbf{x} + \epsilon^2 \lambda + \frac{1}{I} \sum_{i\in[I]} \left[s_i + q^\top \psi_i - \lambda \|\hat{\boldsymbol{\xi}}_i\|_2^2 + \sum_{j\in[N_2+J]} \phi_{ij} q_j^2 \right]$$

subject to

$$\begin{aligned} &\mathbf{x}\in\mathscr{X},\ \lambda\in\mathbb{R}_+,\ s_i\in\mathbb{R},\ \psi_i, \phi_i \in \mathbb{R}^{N_2+J} \quad \forall i\in[I] \\ &\begin{bmatrix} \lambda\mathbb{I} + \boldsymbol{\mathcal{Q}}^\top \operatorname{diag}(\phi_i)\boldsymbol{\mathcal{Q}} & -\frac{1}{2}\boldsymbol{\mathcal{T}}(\mathbf{x})^\top - \boldsymbol{\mathcal{Q}}^\top \operatorname{diag}(\phi_i)\boldsymbol{\mathcal{W}}^\top & -\lambda\hat{\boldsymbol{\xi}}_i - \frac{1}{2}\boldsymbol{\mathcal{Q}}^\top\psi_i \\ -\frac{1}{2}\boldsymbol{\mathcal{T}}(\mathbf{x}) - \boldsymbol{\mathcal{W}}\operatorname{diag}(\phi_i)\boldsymbol{\mathcal{Q}} & \boldsymbol{\mathcal{W}}\operatorname{diag}(\phi_i)\boldsymbol{\mathcal{W}}^\top + \delta\mathbb{I} & \frac{1}{2}(\boldsymbol{\mathcal{W}}\psi_i - h(\mathbf{x})) \\ (-\lambda\hat{\boldsymbol{\xi}}_i - \frac{1}{2}\mathbf{Q}^\top\psi_i)^\top & \frac{1}{2}(\boldsymbol{\mathcal{W}}\psi_i - h(\mathbf{x}))^\top & s_i \end{bmatrix} \succeq_{\mathscr{C}} \mathbf{0} \quad \forall i\in[I]. \end{aligned} \tag{28}$$

Then, the following statements hold.

(i) *If $\delta = 0$ and (1) has complete recourse, then (28) is equivalent to (1).*

(ii) *If $\delta = 0$ and (1) fails to have complete recourse, then (28) provides an upper bound on (1).*

(iii) *If $\delta > 0$, then (28) provides a lower bound on (1).*

(iv) *If $\mathscr{X}$ is compact, then the optimal value of (28) converges to that of (1) for $\delta \downarrow 0$. Moreover, every cluster point $\mathbf{x}^\star$ of a sequence $\{\mathbf{x}_\delta^\star\}_{\delta\downarrow 0}$ of minimizers for (28) is a minimizer for (1).*

Proof. Replacing $\mathcal{Z}(\mathbf{x})$ in (1) with $\bar{\mathcal{Z}}_\delta(\mathbf{x})$ yields (28). Assertion (i) thus follows from Theorem 4, while assertion (ii) follows from Theorem 2, which implies that $\mathcal{Z}(\mathbf{x}) \leqslant \bar{\mathcal{Z}}(\mathbf{x}) = \bar{\mathcal{Z}}_0(\mathbf{x})$ for every $\mathbf{x}\in\mathscr{X}$. Assertion (iii) holds because $\bar{\mathcal{Z}}_\delta(\mathbf{x}) = \underline{\mathcal{Z}}_\delta(\mathbf{x}) \leqslant \underline{\mathcal{Z}}(\mathbf{x}) = \mathcal{Z}(\mathbf{x})$, where the first equality follows from Proposition 2, the inequality holds because problem (25) constitutes a relaxation of (10), and the second equality is due to Theorem 3. As for assertion (iv), recall that $\underline{\mathcal{Z}}_\delta(\mathbf{x})$ is the optimal value of problem (25) and thus constitutes a pointwise supremum of affine functions in $\mathbf{x}$. Therefore, $\bar{\mathcal{Z}}_\delta(\mathbf{x}) = \underline{\mathcal{Z}}_\delta(\mathbf{x})$ is convex and lower semicontinuous in $\mathbf{x}$ for every fixed $\delta > 0$. As $\mathscr{X}$ is compact, we may thus conclude that there exists a minimizer $\mathbf{x}_\delta^\star \in \operatorname{argmin}_{\mathbf{x}\in\mathscr{X}}\{\mathbf{c}^\top\mathbf{x} + \bar{\mathcal{Z}}_\delta(\mathbf{x})\}$ for every $\delta > 0$. Next, note that the sequence of functions $\{\bar{\mathcal{Z}}_\delta(\mathbf{x})\}_{\delta\downarrow 0}$ is nondecreasing and thus epi-converges to $\mathcal{Z}(\mathbf{x})$ by Rockafellar and Wets (2009, proposition 7.4(d)). Moreover, the sequence of minimizers $\{\mathbf{x}_\delta^\star\}_{\delta\downarrow 0}$ admits at least one cluster point $\mathbf{x}^\star \in \mathscr{X}$. By Shapiro et al. (2014, proposition 7.30), $\mathbf{x}^\star$ constitutes a minimizer for (1), and we have

$$\lim_{\delta\downarrow 0} \min_{\mathbf{x}\in\mathscr{X}} \{\mathbf{c}^\top\mathbf{x} + \bar{\mathcal{Z}}_\delta(\mathbf{x})\} = \min_{\mathbf{x}\in\mathscr{X}} \{\mathbf{c}^\top\mathbf{x} + \mathcal{Z}(\mathbf{x})\}.$$

This completes the proof.

Theorem 5 immediately extends to two-stage robust optimization problems of the form (7).

Corollary 1 (Two-Stage Robust Optimization). *Assume that Ξ is bounded and set $I = 1$. Moreover, choose $\epsilon \geqslant 0$ and $\hat{\boldsymbol{\xi}}_1 \in \Xi$ such that $d(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_1) \leqslant \epsilon\ \forall \boldsymbol{\xi} \in \Xi$. Then, the following statements hold.*

(i) *If $\delta = 0$ and (7) has complete recourse, then (28) is equivalent to (7).*

(ii) *If $\delta = 0$ and (7) fails to have complete recourse, then (28) provides an upper bound on (7).*

(iii) *If $\delta > 0$, then (28) provides a lower bound on (7).*

(iv) *If $\mathscr{X}$ is compact, then the optimal value of (28) converges to that of (7) for $\delta \downarrow 0$. Moreover, every cluster point $\mathbf{x}^\star$ of a sequence $\{\mathbf{x}_\delta^\star\}_{\delta\downarrow 0}$ of minimizers for (28) is a minimizer for (7).*

Proof. By construction of ϵ and $\hat{\boldsymbol{\xi}}_1$, the Wasserstein ball $\mathbb{B}_\epsilon^2(\hat{\mathbb{P}}_I)$ contains all Dirac distributions $\delta_{\boldsymbol{\xi}}$, $\boldsymbol{\xi} \in \Xi$. Therefore, the worst-case expected cost (2) reduces to $\max_{\boldsymbol{\xi}\in\Xi} Z(\mathbf{x}, \boldsymbol{\xi})$. The claim thus follows immediately from Theorem 5. □

To our best knowledge, Corollary 1 provides the first exact finite conic programming reformulation for the *generic* two-stage robust optimization problem (7). When the uncertainty appears only in the constraints of the recourse problem (3), approximation schemes based on the cutting plane method are available (Thiele et al. 2009, Zeng and Zhao 2013). These approximations construct increasingly tight lower bounds for the wait-and-see cost $\max_{\boldsymbol{\xi}\in\Xi} Z(\mathbf{x}, \boldsymbol{\xi})$. Each iteration is costly, however, as it involves the solution of a bilinear

maximization problem. Thus, additional assumptions on the uncertainty set Ξ are needed to alleviate the computational burden of generating the cuts. For example, when Ξ is a finite or a budgeted uncertainty set, then one can reformulate the corresponding bilinear maximization problems as mixed-integer linear programs of moderate sizes. If problem (7) fails to have relatively complete recourse, however, then many expensive iterations may be required to obtain a first feasible solution. The results portrayed in Theorem 5 and Corollary 1 motivate an alternative conservative approximation scheme to solve problem (7) that can immediately produce a feasible solution. This is achieved by employing a tractable inner approximation for the copositive cone $\mathscr{C}$. We discuss this approach in the next section.

3.4. A Hierarchy of Semidefinite Programming Approximations

To justify the use of an approximation scheme, we first establish that two-stage distributionally robust linear programs of the form (1) are intractable.

Proposition 3. *The two-stage distributionally robust linear program* (1) *is NP-hard even if* $\mathbf{Q} = \mathbf{0}$ *and* $\mathscr{X}$ *is a polyhedron specified by a list of linear inequalities.*

Proof. Recall from Remark 1 that the distributionally robust linear programs (1) encapsulate the class of all two-stage robust optimization problems of the form (7). The claim thus follows immediately from Guslitser (2002, theorem 3.5), which asserts that two-stage robust optimization problems with fixed recourse are NP-hard. □

The complexity result of Proposition 3 is also plausible in view of Theorem 5 and the known fact that linear programs over copositive cones are generically intractable (Murty and Kabadi 1987). A tractable conservative approximation for (28) can be obtained by replacing the copositive cone $\mathscr{C}$ with

$$\mathscr{C}^0 = \{\mathbf{M} \in \mathbb{S}^K : \mathbf{M} = \mathbf{P} + \mathbf{N},\ \mathbf{P} \geq \mathbf{0},\ \mathbf{N} \geqslant \mathbf{0}\}.$$

By construction, we have $\mathscr{C}^0 \subseteq \mathscr{C}$, but for dimensions $K \leqslant 4$, one can prove that $\mathscr{C}^0 = \mathscr{C}$ (Diananda 1962). For $K > 4$, $\mathscr{C}^0$ is a strict subset of $\mathscr{C}$. In this case, there exists a hierarchy of semidefinite representable cones $\{\mathscr{C}^l\}_{l \geqslant 1}$ that provide increasingly tight inner approximations for $\mathscr{C}$ and converge in finitely many iterations to $\mathscr{C}$ (Parrilo 2000, Bomze and de Klerk 2002, de Klerk and Pasechnik 2002, Lasserre 2009). If these tractable cones are used to replace $\mathscr{C}$ in (28), then the sizes of the resulting approximate problems can, however, become prohibitively large for $l > 0$. In practice, we find that replacing the cone $\mathscr{C}$ with $\mathscr{C}^0$ is sufficient to generate solutions that enjoy an acceptable accuracy.

Theorem 4 implies that if problem (1) has complete recourse, then the copositive program (28) with $\delta = 0$ is equivalent to the two-stage distributionally robust linear program (1). In numerical tests, we observe that strong duality between the conic programs (10) and (14) holds also for many problem instances that violate the complete recourse condition. In all these cases, near-optimal solutions for (1) can be computed by solving the semidefinite programming approximation obtained by setting $\delta = 0$ and replacing the cone $\mathscr{C}$ in (28) with an inner approximation $\mathscr{C}^l$. On the other hand, if strong duality fails to hold, then feasible candidate solutions for (1) can be computed in a similar manner by solving the semidefinite programming approximation (28) for increasingly small values of $\delta > 0$ until a suitable termination criterion is met.

3.5. Accounting for Risk Aversion

Until now, we have studied two-stage distributionally robust optimization problems with an ambiguity-averse but risk-neutral decision maker in mind. The worst-case expectation (2) is a natural objective criterion for such agents. However, many decision makers are both ambiguity averse *and* risk averse, and they may thus prefer to minimize a worst-case optimized certainty equivalent. In the remainder of this paper, we will argue that the main results of this section naturally extend to this setting. Specifically, we consider nondecreasing, convex, and piecewise affine disutility functions of the form $U(y) = \max_{t \in [T]} \{\alpha_t y + \beta_t\}$, where $\boldsymbol{\alpha} \in \mathbb{R}^T_+$, $\boldsymbol{\alpha} \neq \mathbf{0}$ and $\boldsymbol{\beta} \in \mathbb{R}^T$, and we replace the worst-case expectation (2) with the worst-case optimized certainty equivalent,

$$\mathscr{Z}(\mathbf{x}) = \inf_{\theta \in \mathbb{R}} \left\{ \theta + \sup_{\mathbb{P} \in \hat{\mathscr{P}}} \mathbb{E}_{\mathbb{P}}[U(Z(\mathbf{x}, \tilde{\boldsymbol{\xi}}) - \theta)] \right\}, \qquad (29)$$

corresponding to U. The worst-case optimized certainty equivalent determines an optimized payment schedule of the uncertain wait-and-see cost $Z(\mathbf{x}, \tilde{\boldsymbol{\xi}})$ into a fraction θ that is paid here and now and a remainder $Z(\mathbf{x}, \tilde{\boldsymbol{\xi}}) - \theta$ that is paid after the uncertainty has been observed. Optimized certainty equivalents encapsulate mean-variance and conditional value-at-risk measures as special cases; see Ben-Tal and Teboulle (2007). Similar objective criteria are used in Hanasusanto et al. (2016b) and Natarajan et al. (2011) to model the decision maker's risk aversion.

Corollary 2. *Consider the following family of copositive programs parametrized in* δ:

$$\inf \left\{ \mathbf{c}^\top \mathbf{x} + \theta + \epsilon^2 \lambda + \frac{1}{I} \sum_{i \in [I]} s_i \right\}$$
$$\text{s.t. } \mathbf{x} \in \mathscr{X},\ \theta \in \mathbb{R},\ \lambda \in \mathbb{R}_+,\ s_i, \kappa_{it} \in \mathbb{R},\ \boldsymbol{\psi}_{it}, \boldsymbol{\phi}_{it} \in \mathbb{R}^{N_2 + J}$$
$$\forall i \in [I],\ \forall t \in [T],$$

$$\begin{bmatrix} \lambda\mathbb{I}+\boldsymbol{\mathcal{Q}}^\top \operatorname{diag}(\phi_{it})\boldsymbol{\mathcal{Q}} & -\frac{1}{2}\alpha_t\boldsymbol{\mathcal{T}}(\mathbf{x})^\top-\boldsymbol{\mathcal{Q}}^\top\operatorname{diag}(\phi_{it})\boldsymbol{\mathcal{W}}^\top & -\lambda\hat{\boldsymbol{\xi}}_i-\frac{1}{2}\mathbf{Q}^\top\psi_{it} \\ -\frac{1}{2}\alpha_t\boldsymbol{\mathcal{T}}(\mathbf{x})-\boldsymbol{\mathcal{W}}\operatorname{diag}(\phi_{it})\boldsymbol{\mathcal{Q}} & \boldsymbol{\mathcal{W}}\operatorname{diag}(\phi_{it})\boldsymbol{\mathcal{W}}^\top+\delta\mathbb{I} & \frac{1}{2}(\boldsymbol{\mathcal{W}}\psi_{it}-\alpha_t h(\mathbf{x})) \\ (-\lambda\hat{\boldsymbol{\xi}}_i-\frac{1}{2}\mathbf{Q}^\top\psi_{it})^\top & \frac{1}{2}(\boldsymbol{\mathcal{W}}\psi_{it}-\alpha_t h(\mathbf{x}))^\top & s_i+\kappa_{it} \end{bmatrix} \succeq_{\mathscr{C}} \mathbf{0} \quad \forall i\in[I]\,\forall t\in[T],$$

$$\kappa_{it}=\alpha_t\theta-\beta_t-q^\top\psi_{it}+\lambda\|\hat{\boldsymbol{\xi}}_i\|_2^2-\sum_{j\in[N_2+J]}\phi_{itj}q_j^2 \quad \forall i\in[I]\,\forall t\in[T]. \tag{30}$$

If $\mathscr{Z}(\mathbf{x})$ denotes the worst-case expected disutility function (29), then the following statements hold.

(i) *If $\delta=0$ and (1) has complete recourse, then (30) is equivalent to (1).*

(ii) *If $\delta=0$ and (1) fails to have complete recourse, then (30) provides an upper bound on (1).*

(iii) *If $\delta>0$, then (30) provides a lower bound on (1).*

(iv) *If $\mathscr{X}$ is compact, then the optimal value of (30) converges to that of (1) for $\delta\downarrow 0$. Moreover, every cluster point $\mathbf{x}^\star$ of a sequence $\{\mathbf{x}_\delta^\star\}_{\delta\downarrow 0}$ of minimizers for (30) is a minimizer for (1).*

Proof. This is an immediate generalization of Theorem 5. Details are omitted for brevity. □

4. Linear Programming Reformulation for $\mathbf{Q}=\mathbf{0}$

We assume now that the uncertainty affects only the constraints of the recourse problem (3); that is, we assume that $\mathbf{Q}=\mathbf{0}$. Unless stated otherwise, we further assume throughout this section that $\Xi=\mathbb{R}^K$ and that the ambiguity set is constructed using the 1-Wasserstein metric with reference distance $d(\xi_1,\xi_2)=\|\xi_1-\xi_2\|$, where the norm $\|\cdot\|$ is defined through

$$\|\xi\|=\mathbf{e}^\top\max\{w_+\cdot\xi,-w_-\cdot\xi\} \tag{31}$$

for some positive scaling parameters w_+ and w_-. Note that (31) reduces to the 1-norm if $w_+=w_-=1$.

Finally, we always assume that (1) has sufficiently expensive recourse. Under these assumptions, the two-stage distributionally robust linear program (1) admits an equivalent reformulation as a tractable linear program.

4.1. Tractable Formulation

We first establish the main tractability result for the two-stage distributionally robust linear program (1).

Theorem 6. *The two-stage distributionally robust optimization problem (1) is equivalent to the tractable linear program:*

$$\text{minimize}\ \left\{\epsilon\lambda+\frac{1}{I}\sum_{i\in[I]}\mathbf{q}^\top\mathbf{y}_i\right\}$$

subject to

$$\left.\begin{array}{l} \mathbf{x}\in\mathscr{X},\ \lambda\in\mathbb{R}_+,\ \mathbf{y}_i\in\mathbb{R}^{N_2} \quad \forall i\in[I], \\ \phi_k,\ \psi_k\in\mathbb{R}^{N_2} \quad \forall k\in[K], \\ \mathbf{T}(\mathbf{x})\hat{\boldsymbol{\xi}}_i+\mathbf{h}(\mathbf{x})\leqslant\mathbf{W}\mathbf{y}_i \quad \forall i\in[I], \\ \mathbf{q}^\top\phi_k\leqslant\lambda,\ \mathbf{q}^\top\psi_k\leqslant\lambda \\ \mathbf{T}(\mathbf{x})\mathbf{e}_k/w_+\leqslant\mathbf{W}\phi_k,\ -\mathbf{T}(\mathbf{x})\mathbf{e}_k/w_-\leqslant\mathbf{W}\psi_k \end{array}\right\} \quad \forall k\in[K]. \tag{32}$$

Proof. By strong linear programming duality, which holds because problem (1) has sufficiently expensive recourse, we have $Z(\mathbf{x},\xi)=Z_{\mathrm{d}}(\mathbf{x},\xi)$ for every $\mathbf{x}\in\mathscr{X}$ and $\xi\in\mathbb{R}^K$. Theorem 1 thus implies that

$$\mathscr{Z}(\mathbf{x})=\inf_{\lambda\geqslant 0}\epsilon\lambda+\frac{1}{N}\sum_{i\in[I]}\sup_{\xi}\sup_{\substack{\mathbf{p}\geqslant\mathbf{0}\\ \mathbf{W}^\top\mathbf{p}=\mathbf{q}}}(\mathbf{T}(\mathbf{x})\xi+\mathbf{h}(\mathbf{x}))^\top\mathbf{p}-\lambda\|\xi-\hat{\boldsymbol{\xi}}_i\|.$$

Invoking the definition of the dual norm and interchanging the order of the supremum operators over ξ and $\mathbf{p}$, we further obtain

$$\mathscr{Z}(\mathbf{x})=\inf_{\lambda\geqslant 0}\left\{\epsilon\lambda+\frac{1}{N}\sum_{i\in[I]}\sup_{\substack{\mathbf{p}\geqslant\mathbf{0}\\ \mathbf{W}^\top\mathbf{p}=\mathbf{q}}}\left[\sup_{\xi}\left(\inf_{\|\gamma\|_*\leqslant\lambda}(\mathbf{T}(\mathbf{x})\xi+\mathbf{h}(\mathbf{x}))^\top\cdot\mathbf{p}-\gamma^\top\xi+\gamma^\top\hat{\boldsymbol{\xi}}_i\right)\right]\right\}.$$

Next, we interchange the order of the innermost supremum over ξ and the infimum over γ, which is allowed by the classical minimax theorem (Bertsekas 2009, proposition 5.5.4) since γ ranges over a compact set. This yields

$$\mathscr{Z}(\mathbf{x})=\inf_{\lambda\geqslant 0}\left\{\epsilon\lambda+\frac{1}{I}\sum_{i\in[I]}\sup_{\substack{\mathbf{p}\geqslant\mathbf{0}\\ \mathbf{W}^\top\mathbf{p}=\mathbf{q}}}\inf_{\|\gamma\|_*\leqslant\lambda}\left[\sup_{\xi}((\mathbf{T}(\mathbf{x})\xi+\mathbf{h}(\mathbf{x}))^\top\cdot\mathbf{p}-\gamma^\top\xi+\gamma^\top\hat{\boldsymbol{\xi}}_i)\right]\right\}.$$

Evaluating the inner maximization over ξ analytically further yields

$$\begin{aligned}\mathscr{Z}(\mathbf{x})&=\inf_{\lambda\geqslant 0}\left\{\epsilon\lambda+\frac{1}{I}\sum_{i\in[I]}\sup_{\substack{\mathbf{p}\geqslant\mathbf{0}\\ \mathbf{W}^\top\mathbf{p}=\mathbf{q}}}\left[\inf_{\|\gamma\|_*\leqslant\lambda}(\mathbf{h}(\mathbf{x})^\top\mathbf{p}+\gamma^\top\hat{\boldsymbol{\xi}}_i+\chi_{\{\gamma=\mathbf{T}(\mathbf{x})^\top\mathbf{p}\}}(\gamma,\mathbf{p}))\right]\right\}\\ &=\inf_{\lambda\geqslant 0}\left\{\epsilon\lambda+\frac{1}{I}\sum_{i\in[I]}\sup_{\substack{\mathbf{p}\geqslant\mathbf{0}\\ \mathbf{W}^\top\mathbf{p}=\mathbf{q}}}\left[\mathbf{h}(\mathbf{x})^\top\mathbf{p}+(\mathbf{T}(\mathbf{x})^\top\mathbf{p})^\top\hat{\boldsymbol{\xi}}_i+\chi_{\{\|\mathbf{T}(\mathbf{x})^\top\mathbf{p}\|_*\leqslant\lambda\}}(\mathbf{p})\right]\right\}.\end{aligned}$$

The minimization over λ in the last problem can also be evaluated analytically. In fact, the unique optimal solution is $\lambda^\star=\sup\{\|\mathbf{T}(\mathbf{x})^\top\mathbf{p}\|_*:\mathbf{p}\in\mathbb{R}_+^M,\ \mathbf{W}^\top\mathbf{p}=\mathbf{q}\}$. Note that for any $\lambda<\lambda^\star$, the supremum over $\mathbf{p}$ would be

unbounded, and any $\lambda > \lambda^\star$ would incur an unnecessarily high cost as λ is penalized by ϵ in the objective function. We thus obtain

$$\mathcal{Z}(\mathbf{x}) = \inf \left\{ \epsilon\lambda + \frac{1}{I}\sum_{i\in[I]} \sup_{\substack{\mathbf{p}\geqslant\mathbf{0}\\ \mathbf{W}^\top\mathbf{p}=\mathbf{q}}} (\mathbf{h}(\mathbf{x})^\top\mathbf{p} + (\mathbf{T}(\mathbf{x})^\top\mathbf{p})^\top\hat{\boldsymbol{\xi}}_i) \right\}$$
$$\text{s.t. } \lambda\in\mathbb{R}_+,$$
$$\|\mathbf{T}(\mathbf{x})^\top\mathbf{p}\|_* \leqslant \lambda \quad \forall \mathbf{p}\in\mathbb{R}^M_+ : \mathbf{W}^\top\mathbf{p}=\mathbf{q}. \tag{33}$$

Next, the norm dual to (31) is given by

$$\|\mathbf{z}\|_* = \max_{k\in[K]} \left[\max\left\{ \frac{z_k}{w_+}, -\frac{z_k}{w_-} \right\} \right].$$

Thus, the last constraint in (33) can be decomposed into a system of $\mathcal{O}(K)$ linear constraints as follows:

$$\begin{aligned}
&\|\mathbf{T}(\mathbf{x})^\top\mathbf{p}\|_* \leqslant \lambda, \quad \forall \mathbf{p}\in\mathbb{R}^M_+ : \mathbf{W}^\top\mathbf{p}=\mathbf{q} \\
&\iff \sup_{\substack{\mathbf{p}\geqslant\mathbf{0}\\ \mathbf{W}^\top\mathbf{p}=\mathbf{q}}} \mathbf{e}_k^\top\mathbf{T}(\mathbf{x})^\top\mathbf{p}/w_+ \leqslant \lambda, \ \sup_{\substack{\mathbf{p}\geqslant\mathbf{0}\\ \mathbf{W}^\top\mathbf{p}=\mathbf{q}}} -\mathbf{e}_k^\top\mathbf{T}(\mathbf{x})^\top\mathbf{p}/w_- \leqslant \lambda \\
&\qquad\qquad \forall k\in[K], \\
&\iff \exists \boldsymbol{\phi}_k, \boldsymbol{\psi}_k \in \mathbb{R}^{N_2} : \left.\begin{array}{r} \mathbf{q}^\top\boldsymbol{\phi}_k \leqslant \lambda, \mathbf{q}^\top\boldsymbol{\psi}_k \leqslant \lambda \\ \mathbf{T}(\mathbf{x})\mathbf{e}_k/w_+ \leqslant \mathbf{W}\boldsymbol{\phi}_k, \\ -\mathbf{T}(\mathbf{x})\mathbf{e}_k/w_- \leqslant \mathbf{W}\boldsymbol{\psi}_k \end{array}\right\} \quad \forall k\in[K].
\end{aligned}$$

Here, the second equivalence follows from dualizing the linear programs over $\mathbf{p}$, all of which are feasible because problem (1) has sufficiently expensive recourse. The claim then follows from substituting the last constraint system into (33). □

Example 2 (Regression)**.** Consider the least absolute deviations (LAD) regression problem:

$$\begin{aligned}
&\text{minimize} \ \ \mathbb{E}_{\mathbb{P}}[|\mathbf{x}^\top\tilde{\boldsymbol{\xi}} + x_0 - \tilde{\chi}|] \\
&\text{subject to} \ \ (\mathbf{x}, x_0) \in \mathcal{X}.
\end{aligned}$$

The objective of this problem is to find the slope $\mathbf{x}$ and intercept x_0 of an affine function of the explanatory random variables $\tilde{\boldsymbol{\xi}}$ that tightly approximates the independent variable $\tilde{\chi}$ in terms of the mean absolute deviation. In statistics, however, the data-generating distribution $\mathbb{P}$ of $(\tilde{\boldsymbol{\xi}}, \tilde{\chi})$ is never known. Only the empirical distribution $\hat{\mathbb{P}}_I$ corresponding to a set of I training samples is given. In this case $\mathbb{P}$ is ambiguous, and it may make sense to solve the distributionally robust LAD problem:

$$\begin{aligned}
&\text{minimize} \ \ \sup_{\mathbb{P}\in\hat{\mathcal{P}}} \mathbb{E}_{\mathbb{P}}[|\mathbf{x}^\top\tilde{\boldsymbol{\xi}} + x_0 - \tilde{\chi}|] \\
&\text{subject to} \ \ (\mathbf{x}, x_0) \in \mathcal{X},
\end{aligned}$$

which can be identified as an instance of the two-stage distributionally robust optimization problem (1) with recourse function

$$Z((\mathbf{x}, x_0), (\boldsymbol{\xi}, \chi)) = \min\{y : y\in\mathbb{R}, \ y \geqslant \mathbf{x}^\top\boldsymbol{\xi} + x_0 - \chi, \ y \geqslant \chi - x_0 - \mathbf{x}^\top\boldsymbol{\xi}\}.$$

From Equation (33) in the proof of Theorem 6, it is evident that the distributionally robust LAD problem is equivalent to

$$\begin{aligned}
&\text{minimize} \ \ \left\{ \epsilon\|\mathbf{x}\|_* + \frac{1}{N}\sum_{i\in[I]} |\mathbf{x}^\top\hat{\boldsymbol{\xi}}_i + x_0 - \hat{\chi}_i| \right\} \\
&\text{subject to} \ \ (\mathbf{x}, x_0) \in \mathcal{X}.
\end{aligned}$$

Note that the above formulation holds for arbitrary norms (not just the one defined in (32)). The second term in the objective function represents the empirical LAD loss, while the first term acts as a regularizer for the regression coefficient $\mathbf{x}$. If the reference distance is set to the infinity norm, then we recover the celebrated LASSO regularizer (Tibshirani 1996, Wang et al. 2006). On the other hand, if the reference distance is set to the 1-norm, then we obtain an infinity norm regularizer that has recently been employed in the context of logistic regression (Shafieezadeh Abadeh et al. 2015).

Example 3 (Multitask Learning)**.** We can extend Example 2 to a distributionally robust multitask learning problem (Caruana 1996, Baxter 2000) where several regression problems are to be solved simultaneously:

$$\begin{aligned}
&\text{minimize} \ \ \sup_{\mathbb{P}\in\hat{\mathcal{P}}} \mathbb{E}_{\mathbb{P}}[\|\mathbf{X}\tilde{\boldsymbol{\xi}} + \mathbf{x} - \tilde{\boldsymbol{\chi}}\|_1] \\
&\text{subject to} \ \ (\mathbf{X}, \mathbf{x}) \in \mathcal{X}.
\end{aligned}$$

This model has many applications in marketing (Lenk et al. 1996), healthcare (Zhang et al. 2012), and natural language processing (Collobert and Weston 2008), among others. The distributionally robust multitask learning model still constitutes an instance of problem (1), where the recourse function is now given by

$$Z((\mathbf{X}, \mathbf{x}), (\boldsymbol{\xi}, \boldsymbol{\chi})) = \min\{\mathbf{e}^\top\mathbf{y} : \mathbf{y}\in\mathbb{R}^L, \ \mathbf{y} \geqslant \mathbf{X}\boldsymbol{\xi} + \mathbf{x} - \boldsymbol{\chi}, \ \mathbf{y} \geqslant \boldsymbol{\chi} - \mathbf{x} - \mathbf{X}\boldsymbol{\xi}\}.$$

By Theorem 6, this problem is equivalent to the linear program

$$\begin{aligned}
&\text{minimize} \ \ \left\{ \frac{\epsilon}{\min\{w_+, w_-\}} \max_{k\in[K]} \|\mathbf{X}_{\cdot k}\|_1 \right. \\
&\qquad\qquad \left. + \frac{1}{N}\sum_{i\in[I]} \|\mathbf{X}\hat{\boldsymbol{\xi}}_i + \mathbf{x} - \hat{\boldsymbol{\chi}}_i\|_1 \right\} \\
&\text{subject to} \ \ (\mathbf{X}, \mathbf{x}) \in \mathcal{X}.
\end{aligned}$$

Here, the first term in the objective function acts again as a regularizer for the regression coefficient $\mathbf{X}$, while the second term represents the empirical LAD loss.

4.2. Complexity Analysis

Unfortunately, tractability of the distributionally robust linear program (1) is lost when the reference distance is defined via a p-norm with $p > 1$ even if all other conditions of Theorem 6 remain valid. This can be

shown by using a reduction from the NP-hard Matrix Norm Maximization problem (Steinberg 2005).

> Matrix Norm Maximization
> **Instance.** Given a positive semidefinite matrix $\mathbf{M} \in \mathbb{S}_+^K$.
> **Question.** For a fixed $q \in [1,\infty)$, compute the matrix norm $\|\mathbf{M}\|_{\infty,q} = \max_{\|\mathbf{z}\|_\infty \leqslant 1} \|\mathbf{M}\mathbf{z}\|_q$.

Theorem 7. *Computing the optimal value of (1) is NP-hard whenever the reference distance is set to $d(\xi_1,\xi_2) = \|\xi_1 - \xi_2\|_p$ for any $p > 1$, even if $\mathbf{Q} = \mathbf{0}$, $r = 1$, and $\Xi = \mathbb{R}^K$ and there are no first-stage decisions.*

Proof. Fix $p \in (1,\infty]$ and set $q = p/(p-1) \in [1,\infty)$. For any instance $\mathbf{M} \in \mathbb{S}_+^K$ of the Matrix Norm Maximization problem, construct an instance of the distributionally robust linear program (1) as follows. Set the parameters of the recourse problem (3) to $\mathbf{Q} = \mathbf{0}$, $\mathbf{q} = \mathbf{0}$, $\mathbf{T}(\mathbf{x}) = [\mathbf{M} - \mathbf{M}]^\top$, $\mathbf{h}(\mathbf{x}) = \mathbf{0}$, and $\mathbf{W} = [\mathbb{I}\ \mathbb{I}]^\top$. Moreover, assume that there is only one sample $\hat{\xi}_1 = \mathbf{0}$, and set $\epsilon = 1$. Equation (33) in the proof of Theorem 6 implies that problem (1) is equivalent to

$$\begin{array}{ll} \text{minimize} & \lambda + \mathbf{e}^\top \mathbf{y}_1 \\ \text{subject to} & \lambda \in \mathbb{R}_+,\ \mathbf{y}_1 \in \mathbb{R}^{N_2}, \\ & \mathbf{0} \leqslant \mathbf{y}_1, \\ & \|\mathbf{M}(\mathbf{p}_+ - \mathbf{p}_-)\|_q \leqslant \lambda, \quad \forall \mathbf{p}_+, \mathbf{p}_- \in \mathbb{R}_+^K : \mathbf{p}_+ + \mathbf{p}_- = \mathbf{e}. \end{array}$$

Note that $\mathbf{y}_1 = \mathbf{0}$ at optimality irrespective of λ, and thus the optimal value of this problem coincides with

$$\begin{aligned} &\max\{\|\mathbf{M}(\mathbf{p}_+ - \mathbf{p}_-)\|_q : \mathbf{p}_+, \mathbf{p}_- \in \mathbb{R}_+^K,\ \mathbf{p}_+ + \mathbf{p}_- = \mathbf{e}\} \\ &\quad = \max_{\|\mathbf{z}\|_\infty \leqslant 1} \|\mathbf{M}\mathbf{z}\|_q. \end{aligned}$$

We conclude that computing the optimal value of (1) is at least as hard as solving the NP-hard Matrix Norm Maximization problem. □

5. Numerical Results

We now assess the computational and statistical properties of the two-stage distributionally robust linear programs over 2-Wasserstein balls studied in Section 3. All optimization problems are solved with MOSEK v7 using the YALMIP interface (Löfberg 2004) on an 8-core 3.4 GHz computer with 16 GB RAM.

5.1. Approximation Quality

We first assess the error introduced by approximating the copositive cone $\mathscr{C}$ in (28) with its semidefinite inner approximation $\mathscr{C}^0$. To this end, we study recourse problems of the form

$$\begin{aligned} Z(\mathbf{x},\boldsymbol{\xi}) &= \inf_{\mathbf{y} \in \mathbb{R}_+^{N_2}} \{\mathbf{e}^\top \mathbf{y} : \mathbf{A}\boldsymbol{\xi} - \mathbf{b} \leqslant \mathbf{y}\} \\ &= \sum_{n \in [N_2]} \max\{\mathbf{A}_{n:}^\top \boldsymbol{\xi} - b_n, 0\} \\ &= \max_{\mathbf{l} \in \{0,1\}^{N_2}} (\mathbf{A}\boldsymbol{\xi} - \mathbf{b})^\top \mathbf{l}, \end{aligned} \tag{34}$$

where $\mathbf{A} \in [0,1]^{N_2 \times K}$, $\mathbf{b} \in [0,1]^K$, and the random vector $\tilde{\boldsymbol{\xi}} \in \mathbb{R}^K$ is supported on $\Xi = [0,1]^K$. Note that the wait-and-see cost is independent of $\mathbf{x}$ and representable as a sum of N_2 max functions, which can be expressed as the pointwise maximum of 2^{N_2} affine functions in $\boldsymbol{\xi}$. Recourse problems of this type are hard in the stochastic as well as in the robust setting. Indeed, evaluating the expectation of $Z(\mathbf{x},\tilde{\boldsymbol{\xi}})$ is #P-hard even if $N_2 = 1$ and $\tilde{\boldsymbol{\xi}}$ follows the uniform distribution on Ξ (Hanasusanto et al. 2016a, corollary 1). Similarly, evaluating the worst case of $Z(\mathbf{x},\boldsymbol{\xi})$ over all $\boldsymbol{\xi} \in \Xi$ is strongly NP-hard (Hanasusanto 2015, example 1.1.9). The following proposition shows that the worst-case expectation of $Z(\mathbf{x},\tilde{\boldsymbol{\xi}})$ over all distributions of $\tilde{\boldsymbol{\xi}} \in \Xi$ within a given 2-Wasserstein ball can be expressed as the optimal value of a second-order cone program (SOCP) with $\mathscr{O}(2^{N_2})$ constraints.

Proposition 4. *If $\hat{\mathscr{P}} = \mathscr{B}_\epsilon^2(\hat{\mathbb{P}}_I)$, then the worst-case expectation (2) of the wait-and-see cost (34) amounts to*

$$\begin{aligned} \mathscr{Z}(\mathbf{x}) = \inf\ & \epsilon^2\lambda + \frac{1}{I}\sum_{i\in[I]} s_i \\ \text{s.t.}\ & \lambda \in \mathbb{R}_+, s_i \in \mathbb{R}_+, \boldsymbol{\theta}_{il}, \boldsymbol{\eta}_{il} \in \mathbb{R}_+^K \\ & \qquad \forall i \in [I]\ \forall \mathbf{l} \in \{0,1\}^{N_2}, \\ & s_i + b + \lambda\|\hat{\boldsymbol{\xi}}_i\|_2^2 - \mathbf{e}^\top\boldsymbol{\eta}_{il} \geqslant 0 \\ & \qquad \forall i \in [I]\ \forall \mathbf{l} \in \{0,1\}^{N_2}, \\ & \left\| \begin{bmatrix} A^\top\mathbf{l} + 2\lambda\hat{\boldsymbol{\xi}}_i + \boldsymbol{\theta}_{il} - \boldsymbol{\eta}_{il} \\ s_i + \mathbf{b}^\top\mathbf{l} + \lambda\|\hat{\boldsymbol{\xi}}_i\|_2^2 - \mathbf{e}^\top\boldsymbol{\eta}_{il} - \lambda \end{bmatrix} \right\|_2 \\ & \qquad \leqslant s_i + \mathbf{b}^\top\mathbf{l} + \lambda\|\hat{\boldsymbol{\xi}}_i\|_2^2 - \mathbf{e}^\top\boldsymbol{\eta}_{il} + \lambda \\ & \qquad \forall i \in [I]\ \forall \mathbf{l} \in \{0,1\}^{N_2}. \end{aligned} \tag{35}$$

Proof. By Theorem 1, the worst-case expectation (2) is representable as

$$\begin{aligned} &\mathscr{Z}(\mathbf{x}) \\ &= \inf_{\lambda \in \mathbb{R}_+} \left\{ \epsilon^2\lambda + \frac{1}{I}\sum_{i\in[I]} \max_{\boldsymbol{\xi}\in\Xi} \max_{\mathbf{l}\in\{0,1\}^{N_2}} (\mathbf{l}^\top\mathbf{A}\boldsymbol{\xi} - \mathbf{b}^\top\mathbf{l} - \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i\|_2^2) \right\}. \end{aligned}$$

Thus, by introducing auxiliary epigraphical variables s_i, $i \in [I]$, we can reformulate the above optimization problem as the semi-infinite linear program

$$\begin{aligned} \mathscr{Z}(\mathbf{x}) = \inf\ & \epsilon^2\lambda + \frac{1}{I}\sum_{i\in[I]} s_i \\ \text{s.t.}q\ & \lambda \in \mathbb{R}_+,\ s_i \in \mathbb{R}_+ \quad \forall i \in [I], \\ & \max_{\boldsymbol{\xi}\in\Xi} \mathbf{l}^\top\mathbf{A}\boldsymbol{\xi} - \mathbf{b}^\top\mathbf{l} - \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i\|_2^2 \leqslant s_i \\ & \qquad \forall i \in [I]\ \forall \mathbf{l} \in \{0,1\}^{N_2}. \end{aligned}$$

Strong quadratic programming duality implies that the (i,l)th semi-infinite constraint is satisfied if and only if there exist $\boldsymbol{\theta}_{il}, \boldsymbol{\eta}_{il} \in \mathbb{R}_+^K$ that satisfy the hyperbolic constraint

$$\tfrac{1}{4}\|\mathbf{A}^\top\mathbf{l} + 2\lambda\hat{\boldsymbol{\xi}}_i + \boldsymbol{\theta}_{il} - \boldsymbol{\eta}_{il}\|_2^2 \leqslant \lambda(s_i + \mathbf{b}^\top\mathbf{l} + \lambda\|\hat{\boldsymbol{\xi}}_i\|_2^2 - \mathbf{e}^\top\boldsymbol{\eta}_{il}),$$

which is equivalent to the standard SOCP constraints

$$\lambda \geqslant 0, s_i + b + \lambda\|\hat{\boldsymbol{\xi}}_i\|_2^2 - \mathbf{e}^\top\boldsymbol{\eta}_{i1} \geqslant 0,$$
$$\left\|\begin{bmatrix} A^\top\mathbf{1} + 2\lambda\hat{\boldsymbol{\xi}}_i + \boldsymbol{\theta}_{i1} - \boldsymbol{\eta}_{i1} \\ s_i + \mathbf{b}^\top\mathbf{1} + \lambda\|\hat{\boldsymbol{\xi}}_i\|_2^2 - \mathbf{e}^\top\boldsymbol{\eta}_{i1} - \lambda \end{bmatrix}\right\|_2 \leqslant s_i + \mathbf{b}^\top\mathbf{1} + \lambda\|\hat{\boldsymbol{\xi}}_i\|_2^2 - \mathbf{e}^\top\boldsymbol{\eta}_{i1} + \lambda.$$

Thus, the worst-case expectation (2) indeed coincides with the optimal value of (35). □

Because it is hard to evaluate $\mathcal{Z}(\mathbf{x})$ exactly—as reflected by the exponential size of the SOCP (35)—we now investigate two efficient methods for evaluating $\mathcal{Z}(\mathbf{x})$ approximately: the $\mathcal{C}_0$ approximation of the equivalent copositive program (10) and a state-of-the-art quadratic decision rule approximation. The $\mathcal{C}_0$ approximation is obtained by solving (10) with inputs $\mathbf{S} = \mathbb{I}$, $\mathbf{t} = \mathbf{e}$, $\mathbf{Q} = \mathbf{0}$, $\mathbf{q} = \mathbf{e}$, $\mathbf{T}(\mathbf{x}) = [\mathbf{A}^\top\ \mathbf{0}]^\top$, $\mathbf{h}(\mathbf{x}) = [-\mathbf{b}^\top\ \mathbf{0}^\top]^\top$, and $\mathbf{W} = [\mathbb{I}\ \mathbb{I}]^\top$, while approximating the copositive cone $\mathcal{C}$ with $\mathcal{C}_0$.

To develop a decision rule approximation, we first use (Rockafellar and Wets 2009, theorem 14.60) to reformulate (2) as

$$\mathcal{Z}(\mathbf{x}) = \inf_{\mathbf{y}\in\mathcal{L}_{K,N_2}} \left\{ \sup_{\mathbb{P}\in\hat{\mathcal{P}}} \mathbb{E}_{\mathbb{P}}[\mathbf{e}^\top\mathbf{y}(\tilde{\boldsymbol{\xi}})]\colon \mathbf{A}\boldsymbol{\xi} - \mathbf{b} \leqslant \mathbf{y}(\boldsymbol{\xi})\ \forall\,\boldsymbol{\xi}\in\Xi,\ \mathbf{y}(\boldsymbol{\xi}) \geqslant \mathbf{0}\ \forall\,\boldsymbol{\xi}\in\Xi \right\}, \quad (36)$$

where $\mathcal{L}_{K,N_2}$ denotes the linear space of all measurable functions from $\mathbb{R}^K$ to $\mathbb{R}^{N_2}$. A tractable upper bound on $\mathcal{Z}(\mathbf{x})$ is obtained by restricting $\mathcal{L}_{K,N_2}$ to the subspace of all affine functions; see, for example, Gao and Kleywegt (2016). A tighter tractable upper bound can be obtained, however, by restricting $\mathcal{L}_{K,N_2}$ to the subspace of all quadratic functions and by conservatively approximating the emerging semi-infinite constraints by semidefinite constraints using the approximate $\mathcal{S}$-lemma. Quadratic decision rule approximations of this type are also studied in Hanasusanto et al. (2015).

We run numerical experiments for different values of the uncertainty dimension K and the sample size I, and we set the Wasserstein radius to $\epsilon = 1/\sqrt{I}$, thus enforcing the scaling rule advocated in Blanchet and Kang (2016) and Zhao and Guan (2018).[1] All results are averaged over 100 instances generated randomly as follows. We sample the dimension N_2 of the wait-and-see decision uniformly at random from $\{1, 2, \ldots, \lceil\log(K+1)\rceil\}$, which guarantees that the SOCP (35) grows at most polynomially with K and I. Next, we sample $\mathbf{A}$ uniformly from $[0,1]^{N_2\times K}$ and $\mathbf{b}$ uniformly from $[0, \mathbf{e}^\top\mathbf{A}_{1:}]\times\cdots\times[0, \mathbf{e}^\top\mathbf{A}_{N_2:}]$. We then generate independent training samples $\{\hat{\boldsymbol{\xi}}_i\}_{i\in[I]}$ from the uniform distribution on $[0,1]^K$. Finally, we evaluate the worst-case expectation (2) exactly by solving the SOCP (35), and also approximately by computing the $\mathcal{C}^0$ and the quadratic decision rule approximations.

Table 1 reports the optimality gaps of the two approximations relative to the exact worst-case expectation, averaged across all solvable instances. While the optimality gaps of the $\mathcal{C}^0$ approximation remain consistently below 2.3%, the state-of-the-art quadratic decision rule approximation can incur alarmingly large optimality gaps of more than 100%. On the other hand, while MOSEK is able to solve all instances of the quadratic decision rule approximation, it encounters numerical difficulties when solving some of the larger instances of the $\mathcal{C}^0$ approximation. For $K = 64$, for instance, the underlying semidefinite constraints involve blocks of the size 139×139, which pose a distinct challenge for state-of-the-art interior point solvers. The percentages of all instances of the $\mathcal{C}^0$ approximation that could be solved to global optimality are reported in Table 2. The average runtimes of both approximations are presented in Table 3.

5.2. Out-of-Sample Performance

Next, we assess the out-of-sample performance of different data-driven policies in the context of a multi-item newsvendor problem, where an inventory planner has to select a vector $\mathbf{x} \in \mathbb{R}^K_+$ of order quantities for K different products at the beginning of a sales period. We

Table 1. Optimality Gaps (in Percent) of the $\mathcal{C}^0$ Approximation (Left Column of Each Value of K) and the Quadratic Decision Rule Approximation (Right Column of Each Value of K)

	K													
I	1		2		4		8		16		32		64	
5	0.0	0.8	0.0	2.8	0.0	5.2	0.0	3.9	0.5	6.7	0.3	5.5	0.5	3.0
10	0.0	2.2	0.0	6.1	0.0	11.0	0.0	13.0	0.5	15.0	0.4	16.1	0.8	10.2
20	0.0	4.9	0.0	10.6	0.0	17.6	0.0	20.0	0.5	27.6	0.5	33.1	1.9	19.0
40	0.0	8.6	0.0	14.9	0.0	25.4	0.0	28.2	0.8	55.0	0.6	65.2	1.8	54.0
80	0.0	12.4	0.0	19.9	0.0	33.2	0.0	42.0	0.9	77.1	0.8	117.1	0.0	121.1
160	0.0	18.2	0.0	26.8	0.0	43.1	0.0	51.2	1.3	129.6	1.3	201.2	—	220.8
320	0.0	25.3	0.0	34.4	0.0	58.9	0.0	74.3	2.3	237.7	3.7	254.3	—	416.6
640	0.0	33.4	0.0	43.7	0.0	79.1	0.0	102.3	2.3	343.9	0.2	498.3	—	1,137.1

Table 2. Percentage of Solvable Instances from Using the $\mathscr{C}^0$ Approximation

(I,K)									
(10,64)	(20,64)	(40,64)	(80,64)	(160,32)	(160,64)	(320,32)	(320,64)	(640,32)	(640,64)
95	48	6	1	88	0	19	0	2	0

Note. We report only those (I,K) pairs for which fewer than 100% of all instances were solved to optimality.

Table 3. Solution Times in Seconds of the $\mathscr{C}^0$ Approximation (Left) and the Quadratic Decision Rule Approximation (Right)

	K													
	1		2		4		8		16		32		64	
I														
5	<0.1	<0.1	<0.1	<0.1	0.1	0.1	0.5	0.3	2.3	0.5	41.5	6.2	1,375.9	230.4
10	<0.1	<0.1	0.2	0.1	0.4	0.1	0.8	0.5	5.2	1.0	93.9	12.1	3,091.6	355.1
20	0.1	0.1	0.5	0.1	0.7	0.2	1.9	0.9	11.6	1.4	194.2	26.6	5,799.2	490.2
40	0.3	0.2	1.6	0.3	1.2	0.3	3.0	0.7	25.0	3.0	398.7	52.2	11,276.2	1,617.3
80	0.6	0.3	3.3	0.5	2.2	0.5	8.1	1.1	53.5	6.7	905.9	116.9	19,887.3	3,134.0
160	1.3	0.8	3.2	1.1	2.6	0.5	19.4	2.4	109.9	17.7	1,956.5	253.7	—	7,586.9
320	0.8	2.0	2.1	0.5	12.7	1.0	41.5	5.2	251.4	40.8	3,715.2	415.4	—	16,511.3
640	1.7	0.7	31.3	1.1	15.5	2.1	86.8	15.1	514.7	79.1	9,777.1	1,441.9	—	22,365.2

assume that the total order quantity $\mathbf{e}^\top\mathbf{x}$ may not exceed a given budget B. The demands of the products can be described by a random vector $\tilde{\boldsymbol{\xi}} \in \mathbb{R}^K_+$ that follows an unknown multivariate distribution $\mathbb{P}^\star$. We also assume that there are no ordering costs but that excess inventory of the kth product incurs a per-unit holding cost of b_k, while unmet demand incurs a per-unit stockout cost of s_k. The total cost of an order $\mathbf{x}$ incurred in scenario $\boldsymbol{\xi}$ thus amounts to

$$Z(\mathbf{x},\boldsymbol{\xi}) = \inf_{\mathbf{y}\in\mathbb{R}^K} \{\mathbf{e}^\top\mathbf{y}\colon \operatorname{diag}(\mathbf{b})(\mathbf{x}-\boldsymbol{\xi}) \leqslant \mathbf{y},\ \operatorname{diag}(\mathbf{s})(\boldsymbol{\xi}-\mathbf{x}) \leqslant \mathbf{y}\},$$

where $\mathbf{b} = (b_1,\ldots,b_K)^\top$ and $\mathbf{s} = (s_1,\ldots,s_K)^\top$. By construction, this recourse problem has sufficiently expensive recourse as well as complete recourse. We assume that the inventory planner is both risk averse and ambiguity averse and thus solves the two-stage distributionally robust linear program:

$$\begin{array}{ll} \text{minimize} & \sup\limits_{\mathbb{P}\in\hat{\mathscr{P}}} \mathbb{P}\text{-CVaR}_\rho[Z(\mathbf{x},\tilde{\boldsymbol{\xi}})] \\ \text{subject to} & \mathbf{x}\in\mathbb{R}^K_+, \\ & \mathbf{e}^\top\mathbf{x} \leqslant B, \end{array} \tag{37}$$

which minimizes the worst-case conditional value at risk (CVaR) of $Z(\mathbf{x},\tilde{\boldsymbol{\xi}})$ at level $\rho \in (0,1]$, where the worst case is taken over all distributions within some ambiguity set $\hat{\mathscr{P}}$. For any fixed $\mathbb{P} \in \hat{\mathscr{P}}$, the $\mathbb{P}$-CVaR of $Z(\mathbf{x},\tilde{\boldsymbol{\xi}})$ at level ρ is defined through

$$\mathbb{P}\text{-CVaR}_\rho[Z(\mathbf{x},\tilde{\boldsymbol{\xi}})] = \inf_{\theta\in\mathbb{R}} \left\{\theta + \frac{1}{\rho}\mathbb{E}_\mathbb{P}[\max\{Z(\mathbf{x},\tilde{\boldsymbol{\xi}})-\theta,0\}]\right\}$$

and can be viewed as the conditional expectation of $Z(\mathbf{x},\tilde{\boldsymbol{\xi}})$ above its $(1-\rho)$-percentile under $\mathbb{P}$ (Rockafellar and Uryasev 2000). Note that the worst-case CVaR can be viewed as an instance of the worst-case optimized certainty equivalent (29) corresponding to the disutility function $U(y) = \max\{y,0\}$.

In the following, we review different approaches to construct $\hat{\mathscr{P}}$ from I demand samples $\hat{\boldsymbol{\xi}}_1,\ldots,\hat{\boldsymbol{\xi}}_I$, and we show that in each case problem (37) can be reformulated as a conic program. The first possibility is to set $\hat{\mathscr{P}}$ to the 2-Wasserstein ball $\mathscr{B}^2_\epsilon(\hat{\mathbb{P}}_I)$ around the empirical distribution on the demand samples as in Section 3.

Proposition 5. *If $\hat{\mathscr{P}} = \mathscr{B}^2_\epsilon(\hat{\mathbb{P}}_I)$, then (37) is equivalent to the copositive program*

$$\begin{array}{ll} \text{minimize} & \theta + \dfrac{1}{\rho}\left(\epsilon^2\lambda + \dfrac{1}{I}\displaystyle\sum_{i\in[I]} s_i\right) \\ \text{subject to} & \mathbf{x}\in\mathbb{R}^K_+, \lambda, s_i \in \mathbb{R}_+, \theta\in\mathbb{R}, \boldsymbol{\psi}_i, \boldsymbol{\phi}_i \in \mathbb{R}^{N_2+J}, \quad \forall i\in[I] \\ & \mathbf{e}^\top\mathbf{x} \leqslant B \\ & \begin{bmatrix} \lambda\mathbb{I} & -\frac{1}{2}\mathbf{T}(\mathbf{x})^\top & -\lambda\hat{\boldsymbol{\xi}}_i \\ -\frac{1}{2}\mathbf{T}(\mathbf{x}) & \mathbf{W}\operatorname{diag}(\boldsymbol{\phi}_i)\mathbf{W}^\top & \frac{1}{2}(\mathbf{W}\boldsymbol{\psi}_i - \mathbf{h}(\mathbf{x})) \\ -\lambda\hat{\boldsymbol{\xi}}_i^\top & \frac{1}{2}(\mathbf{W}\boldsymbol{\psi}_i - \mathbf{h}(\mathbf{x}))^\top & s_i + \theta - \mathbf{e}^\top(\boldsymbol{\psi}_i + \boldsymbol{\phi}_i) + \lambda\|\hat{\boldsymbol{\xi}}_i\|_2^2 \end{bmatrix} \succeq_{\mathscr{C}} \mathbf{0} \quad \forall i\in[I], \end{array} \tag{38}$$

where $\mathbf{T}(\mathbf{x}) = [-\operatorname{diag}(\mathbf{b})\ \operatorname{diag}(\mathbf{s})]^\top$, $\mathbf{h}(\mathbf{x}) = [\mathbf{x}^\top\operatorname{diag}(\mathbf{b}), -\mathbf{x}^\top\operatorname{diag}(\mathbf{s})]^\top$, *and* $\mathbf{W} = [\mathbb{I}\ \mathbb{I}]^\top$.

Proof. The claim follows immediately from Corollary 2 and the observation that (37) has sufficiently expensive as well as complete recourse. □

As a second possibility, we can use the I demand samples to estimate the sample mean $\hat{\boldsymbol{\mu}} = (1/I)\sum_{i\in[I]}\hat{\boldsymbol{\xi}}_i$

and the sample covariance matrix $\hat{\mathbf{\Sigma}}=(1/I)\sum_{i\in[I]}(\hat{\boldsymbol{\xi}}_i-\hat{\boldsymbol{\mu}})\cdot(\hat{\boldsymbol{\xi}}_i-\hat{\boldsymbol{\mu}})^\top$ of $\tilde{\boldsymbol{\xi}}$, which can, in turn, be used to construct a Chebyshev ambiguity set of the form

$$\mathscr{P}(\hat{\boldsymbol{\mu}},\hat{\mathbf{\Sigma}},\gamma_1,\gamma_2)=\left\{\mathbb{P}\in\mathscr{M}^2(\mathbb{R}^K_+):\begin{array}{l}(\mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}]-\hat{\boldsymbol{\mu}})^\top\hat{\mathbf{\Sigma}}^{-1}(\mathbb{E}_{\mathbb{P}}[\tilde{\boldsymbol{\xi}}]-\hat{\boldsymbol{\mu}})\leqslant\gamma_1\\ \mathbb{E}_{\mathbb{P}}[(\tilde{\boldsymbol{\xi}}-\hat{\boldsymbol{\mu}})(\tilde{\boldsymbol{\xi}}-\hat{\boldsymbol{\mu}})^\top]\preceq(1+\gamma_2)\hat{\mathbf{\Sigma}}\end{array}\right\}, \tag{39}$$

where $\gamma_1,\gamma_2\in\mathbb{R}_+$ represent two confidence parameters. This ambiguity set has been proposed in Delage and Ye (2010).

Proposition 6. *If $\hat{\mathscr{P}}=\mathscr{P}(\hat{\boldsymbol{\mu}},\hat{\mathbf{\Sigma}},\gamma_1,\gamma_2)$, then (37) is equivalent to the copositive program*

$$\begin{array}{ll}\text{minimize} & \{\theta+(1/\rho)[s+\mathrm{tr}((\gamma_2\hat{\mathbf{\Sigma}}+\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top)\mathbf{M})+\hat{\boldsymbol{\mu}}^\top\mathbf{m}+\sqrt{\gamma_1}\|\hat{\mathbf{\Sigma}}^{1/2}(\mathbf{m}+2\mathbf{M}\hat{\boldsymbol{\mu}})\|_2]\}\\ \text{subject to} & \mathbf{x}\in\mathbb{R}^K_+,\ \theta,s\in\mathbb{R},\ \mathbf{m},\boldsymbol{\psi},\boldsymbol{\phi}\in\mathbb{R}^K,\ \mathbf{M}\in\mathbb{S}^K_+,\\ & \mathbf{e}^\top\mathbf{x}\leqslant B,\\ & \begin{bmatrix}\mathbf{M} & -\frac{1}{2}\mathbf{T}(\mathbf{x})^\top & \frac{1}{2}\mathbf{m}\\ -\frac{1}{2}\mathbf{T}(\mathbf{x}) & \mathbf{W}\,\mathrm{diag}(\boldsymbol{\phi})\mathbf{W}^\top & \frac{1}{2}(\mathbf{W}\boldsymbol{\psi}-\mathbf{h}(\mathbf{x}))\\ \frac{1}{2}\mathbf{m}^\top & \frac{1}{2}(\mathbf{W}\boldsymbol{\psi}-\mathbf{h}(\mathbf{x}))^\top & s+\theta-\mathbf{e}^\top(\boldsymbol{\psi}+\boldsymbol{\phi})\end{bmatrix}\succeq_{\mathscr{C}}\mathbf{0},\\ & \begin{bmatrix}\mathbf{M} & \frac{1}{2}\mathbf{m}\\ \frac{1}{2}\mathbf{m}^\top & s\end{bmatrix}\succeq_{\mathscr{C}}\mathbf{0},\end{array} \tag{40}$$

where $\mathbf{T}(\mathbf{x})=[-\mathrm{diag}(\mathbf{b})\ \mathrm{diag}(\mathbf{s})]^\top$, $\mathbf{h}(\mathbf{x})=[\mathbf{x}^\top\mathrm{diag}(\mathbf{b}),-\mathbf{x}^\top\mathrm{diag}(\mathbf{s})]^\top$, *and* $\mathbf{W}=[\mathbb{I}\ \mathbb{I}]^\top$.

Proof. As in the proof of Proposition 5, we may transform (37) to a worst-case expected disutility minimization problem by using Sion's minimax theorem (Sion 1958). The worst-case expectation in the resulting objective function can then be reexpressed as a maximization problem over completely positive cones by leveraging ideas from Natarajan et al. (2011, section 4.4). Finally, problem (40) is obtained via strong conic duality, which allows us to convert the completely positive maximization problem into an equivalent copositive minimization problem. □

Distributionally robust multi-item newsvendor problems with Chebyshev ambiguity sets are known to be NP-hard even if $\gamma_1=\gamma_2=0$ and $\tilde{\boldsymbol{\xi}}$ is supported on $\mathbb{R}^K$, but they admit tractable conservative approximations based on quadratic decision rules (Hanasusanto et al. 2015).

A third possibility is to set $\hat{\mathscr{P}}=\{\hat{\mathbb{P}}_I\}$. This singleton ambiguity set corresponds to a Wasserstein ball around the empirical distribution with radius $\epsilon=0$. Problem (37) then simply reduces to the corresponding sample average approximation (SAA) problem, which is equivalent to a tractable linear program.

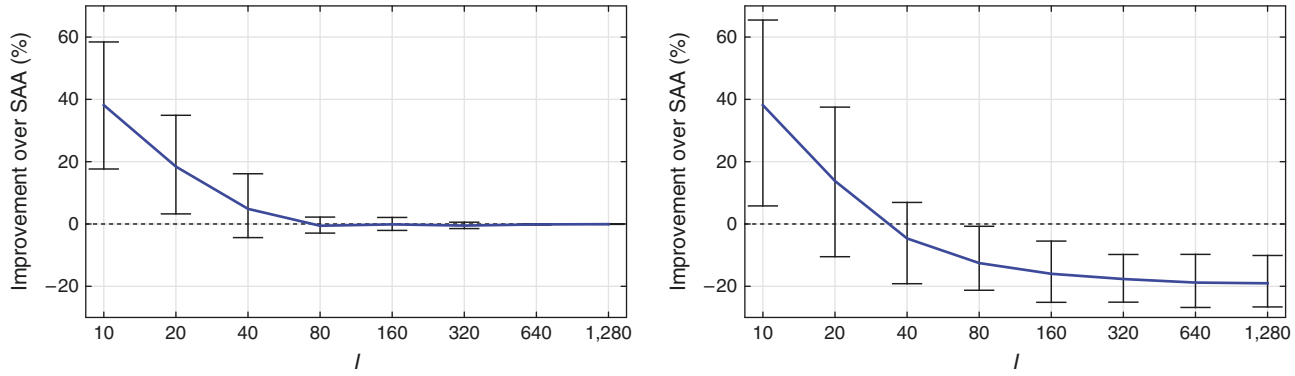
To assess the performance of the Wasserstein, Chebyshev, and SAA policies obtained from the respective distributionally robust optimization models, we conduct out-of-sample experiments for the K-item newsvendor problem with $K=3$ and training data sets containing $I=10$, 20, 40, 80, 160, 320, 640, and 1,028 independent samples. In all experiments, we replace each copositive cone $\mathscr{C}$ appearing in (38) and (40) with its first inner approximation $\mathscr{C}^0$. We fix the vectors of holding and stockout costs to $\mathbf{b}=\mathbf{e}$ and $\mathbf{s}=10\mathbf{e}$, respectively, and we set the ordering budget to $B=30$. We further fix the risk level of the CVaR to $\rho=10\%$.

The results of all experiments are averaged over 100 random trials generated in the following manner. The true demand distribution $\mathbb{P}^\star$ of $\tilde{\boldsymbol{\xi}}$ is assumed to be lognormal; that is, $\tilde{\xi}_k=\exp(\tilde{\chi}_k)$, $k\in[K]$, where the $\tilde{\chi}_k$, $k\in[K]$, represent jointly normally distributed random variables with first- and second-order moments given by $\boldsymbol{\nu}\in\mathbb{R}^K_+$ and $\mathbf{\Sigma}\in\mathbb{S}^K_+$, respectively. In each trial, we sample $\boldsymbol{\nu}$ uniformly at random from $[0,2]^K$ while the matrix $\mathbf{\Sigma}$ is generated randomly using the following procedure. We set the vector of standard deviations to $\boldsymbol{\sigma}=(1/4)\mathbf{e}$, sample a random correlation matrix $\mathbf{C}\in\mathbb{S}^K_+$ using the MATLAB command "`gallery('randcorr',3)`," and set $\mathbf{\Sigma}=\mathrm{diag}(\boldsymbol{\sigma})\mathbf{C}\,\mathrm{diag}(\boldsymbol{\sigma})+\boldsymbol{\nu}\boldsymbol{\nu}^\top$. Next, we sample I independent training samples $\{\hat{\boldsymbol{\xi}}_i\}_{i\in[I]}$ from $\mathbb{P}^\star$. We then compute the Wasserstein policy $\mathbf{x}^\star_{\mathrm{Wass}}$ by solving (38) with $\mathscr{C}^0$ instead of $\mathscr{C}$ and where the Wasserstein radius ϵ is chosen by fivefold cross-validation so as to minimize the out-of-sample risk (Efron and Tibshirani 1994, Hastie et al. 2001). Similarly, the Chebhyshev policy $\mathbf{x}^\star_{\mathrm{Cheb}}$ is obtained by solving (40) with $\mathscr{C}^0$ instead of $\mathscr{C}$ and where the confidence parameters are again determined via fivefold cross-validation. Finally, we compute the SAA policy $\mathbf{x}^\star_{\mathrm{SAA}}$ by solving (37) with $\hat{\mathscr{P}}=\{\hat{\mathbb{P}}_I\}$. The out-of-sample risk $\mathbb{P}^\star$-$\mathrm{CVaR}_\rho[Z(\mathbf{x}^\star,\tilde{\boldsymbol{\xi}})]$ of each of the three data-driven strategies $\mathbf{x}^\star_{\mathrm{Wass}}$, $\mathbf{x}^\star_{\mathrm{Cheb}}$, and $\mathbf{x}^\star_{\mathrm{SAA}}$ is then estimated at high accuracy using 20,000 test samples from $\mathbb{P}^\star$.

Figure 1 visualizes the out-of-sample risk of the Wasserstein and Chebyshev policies relative to the SAA policy as a function of the training sample size I. Observe that the Wasserstein policy dominates the SAA policy with high confidence uniformly across all sample sizes. Moreover, for training data sets of size $I\leqslant 20$, both the Wasserstein and Chebyshev policies outperform the SAA policy with high confidence by more than 20%. This suggests that the distributionally robust policies are preferable whenever there is significant ambiguity about the true distribution $\mathbb{P}^\star$. While the Wasserstein policy consistently outperforms the SAA policy, the quality of the Chebyshev policy starts to deteriorate for $I\geqslant 30$.

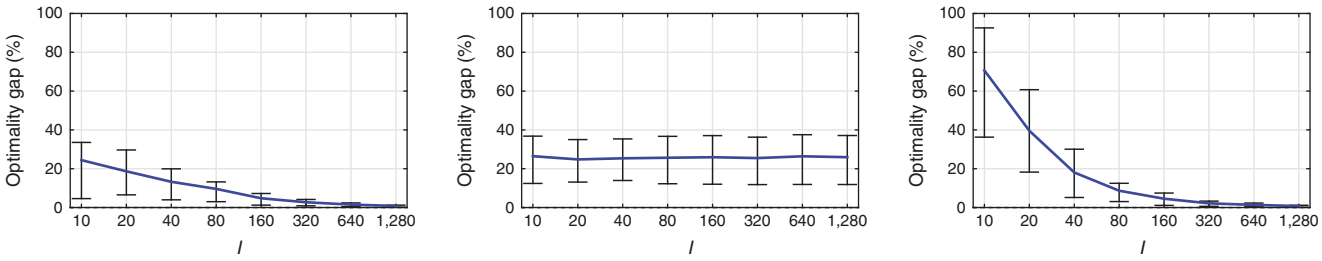
Figure 2 depicts the optimality gaps of the three policies with respect to the true optimal policy, which we estimate by solving another SAA problem using 20,000

Figure 1. (Color online) Improvement of the Wasserstein (Left) and Chebyshev (Right) Policy Relative to the SAA Policy in Terms of Out-of-Sample CVaR



Note. The solid blue lines represent the mean, and the error bars visualize the 20% and 80% quantiles of the relative improvement, respectively.

Figure 2. (Color online) Optimality Gaps of the Wasserstein (Left), Chebyshev (Middle), and SAA (Right) Policies



Note. The solid blue lines represent the mean, and the error bars visualize the 20% and 80% quantiles of the optimality gaps, respectively.

samples from $\mathbb{P}^\star$. For $I = 10$, we find that the SAA policy is ~70% suboptimal, while both the Chebyshev and Wasserstein policies are only ~25% suboptimal on average. For $I = 20$, the SAA policy remains ~40% suboptimal, while the Wassertein policy exhibits a marginally better suboptimality of ~20% on average. The Chebyshev policy, on the other hand, reaches a steady state already for $I = 10$ with an average suboptimality of about 25%. This suggests that the empirical estimates of the first- and second-order moments are already accurate enough for small training data sets. Unfortunately, as the sample size grows, the Chebyshev policy cannot improve, as the first two moments are insufficient to describe the entire shape of the true distribution $\mathbb{P}^\star$. The Wasserstein policy, on the other hand, strikes a good balance between robustness and asymptotic consistency. In particular, we find that it converges quickly to the true optimal policy as the size of the training data set grows.

Acknowledgments

The authors gratefully acknowledge the two anonymous reviewers whose comments led to substantial improvements of the paper.

Endnote

[1] We also ran all experiments with $\epsilon = 1$ and $\epsilon = 1/I$ but did not observe any qualitative changes in the results.

References

Ardestani-Jaafari A, Delage E (2016) Linearized robust counterparts of two-stage robust optimization problems with applications in operations management. Working paper, HEC Montreal, Montreal.

Baxter J (2000) A model of inductive bias learning. *J. Artificial Intelligence Res.* 12(3):149–198.

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

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton University Press, Princeton, NJ).

Ben-Tal A, Goryashko A, Guslitzer E, Nemirovski A (2004) Adjustable robust solutions of uncertain linear programs. *Math. Programming A* 99(2):351–376.

Bertsekas DP (2009) *Convex Optimization Theory* (Athena Scientific, Nashua, NH).

Bertsimas D, Gupta V, Kallus N (2017) Robust sample average approximation. *Math. Programming*, ePub ahead of June 29, https://doi.org/10.1007/s10107-017-1174-z.

Bertsimas D, Doan XV, Natarajan K, Teo C-P (2010) Models for minimax stochastic linear optimization problems with risk aversion. *Math. Oper. Res.* 35(3):580–602.

Blanchet J, Kang Y (2016) Sample out-of-sample inference based on Wasserstein distance. Working paper, Columbia University, New York.

Bomze IM, de Klerk E (2002) Solving standard quadratic optimization problems via linear, semidefinite and copositive programming. *J. Global Optim.* 24(2):163–185.

Burer S (2009) On the copositive representation of binary and continuous nonconvex quadratic programs. *Math. Programming A* 120(2):479–495.

Caruana R (1996) Multi-task learning. *Machine Learn.* 28(1):41–75.

Collobert R, Weston J (2008) A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proc. 25th Internat. Conf. Machine Learn.* (ACM, New York), 160–167.

de Klerk E, Pasechnik DV (2002) Approximation of the stability number of a graph via copositive programming. *SIAM J. Optim.* 12(4): 875–892.

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612.

Diananda PH (1962) On non-negative forms in real variables some or all of which are non-negative. *Math. Proc. Cambridge Philos. Soc.* 58(1):17–25.

Dupačová J (1966) On minimax solutions of stochastic linear programming problems. *Časopis pro pěston matematiky* 91(4): 423–430. [Also cited as Žáčková J.]

Efron B, Tibshirani RJ (1994) *An Introduction to the Bootstrap* (CRC Press, Boca Raton, FL).

Gao R, Kleywegt AJ (2016) Distributionally robust stochastic optimization with Wasserstein distance. Working paper, Georgia Institue of Technology, Atlanta.

Georghiou A, Wiesemann W, Kuhn D (2015) Generalized decision rule approximations for stochastic programming via liftings. *Math. Programming A* 152(1–2):301–338.

Gilboa I, Schmeidler D (1989) Maxmin expected utility with non-unique prior. *J. Math. Econom.* 18(2):141–153.

Goh J, Sim M (2010) Distributionally robust optimization and its tractable approximations. *Oper. Res.* 58(4, Part 1):902–917.

Guslitser E (2002) Uncertainty-immunized solutions in linear programming. Unpublished master's thesis, Technion–Israel Institute of Technology, Haifa.

Hadjiyiannis MJ, Goulart PJ, Kuhn D (2011) A scenario approach for estimating the suboptimality of linear decision rules in two-stage robust optimization. *IEEE Conf. Decision Control Eur. Control Conf.*, 7386–7391.

Hanasusanto GA (2015) Decision making under uncertainty: Robust and data-driven approaches. Unpublished Ph.D. thesis, Imperial College London, London.

Hanasusanto GA, Kuhn D, Wiesemann W (2016a) A comment on "computational complexity of stochastic programming problems." *Math. Programming A* 159(1):557–569.

Hanasusanto GA, Kuhn D, Wiesemann W (2016b) *K*-adaptability in two-stage distributionally robust binary programming. *Oper. Res. Lett.* 44(1):6–11.

Hanasusanto GA, Kuhn D, Wallace SW, Zymler S (2015) Distributionally robust multi-item newsvendor problems with multimodal demand distributions. *Math. Programming A* 152(1–2):1–32.

Hastie T, Tibshirani R, Friedman J (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York).

Jiang R, Guan Y (2018) Risk-averse two-stage stochastic program with distributional ambiguity. *Oper. Res*. Forthcoming.

Kleywegt AJ, Shapiro A, de Mello TH (2002) The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.* 12(2):479–502.

Kong Q, Lee C-Y, Teo C-P, Zheng Z (2013) Scheduling arrivals to a stochastic service delivery system using copositive cones. *Oper. Res.* 61(3):711–726.

Lasserre JB (2009) Convexity in semialgebraic geometry and polynomial optimization. *SIAM J. Optim.* 19(4):1995–2014.

Lenk PJ, DeSarbo WS, Green PE, Young RM (1996) Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Sci.* 15(2):173–191.

Li X, Natarajan K, Teo C-P, Zheng Z (2014) Distributionally robust mixed integer linear programs: Persistency models with applications. *Eur. J. Oper. Res.* 233(3):459–473.

Löfberg J (2004) YALMIP: A toolbox for modeling and optimization in MATLAB. *IEEE Internat. Sympos. Comput. Aided Control Systems Design*, 284–289.

Love D, Bayraksan G (2016) Phi-divergence constrained ambiguous stochastic programs for data-driven optimization. Working paper, Ohio State University, Columbus.

Mohajerin Esfahani P, Kuhn D (2017) Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Programming*, ePub ahead of print July 7, https://doi.org/10.1007/s10107-017-1172-1.

Murty KG, Kabadi SN (1987) Some NP-complete problems in quadratic and nonlinear programming. *Math. Programming* 39(2):117–129.

Natarajan K, Teo C-P (2017) On reduced semidefinite programs for second order moment bounds with applications. *Math. Programming A* 161(1–2):487–518.

Natarajan K, Teo C-P, Zheng Z (2011) Mixed 0-1 linear programs under objective uncertainty: A completely positive representation. *Oper. Res.* 59(3):713–728.

Parrilo PA (2000) Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization. Unpublished Ph.D. thesis, California Institute of Technology, Pasadena.

Pflug G, Pichler A (2014) *Multistage Stochastic Optimization* (Springer International, Cham, Switzerland).

Pflug G, Wozabal D (2007) Ambiguity in portfolio selection. *Quant. Finance* 7(4):435–442.

Rockafellar RT, Uryasev S (2000) Optimization of conditional value-at-risk. *J. Risk* 2(3):21–42.

Rockafellar RT, Wets RJ-B (2009) *Variational Analysis* (Springer-Verlag, Berlin).

Scarf HE (1958) A min-max solution to an inventory problem. Arrow KJ, Karlin S, Scarf HE, eds. *Studies in Mathematical Theory of Inventory and Production* (Stanford University Press, Stanford, CA), 201–209.

Shafieezadeh Abadeh S, Mohajerin Esfahani P, Kuhn D (2015) Distributionally robust logistic regression. *Adv. Neural Inform. Processing Systems*, 1576–1584.

Shaked-Monderer N, Berman A, Bomze IM, Jarre F, Schachinger W (2015) New results on the cp-rank and related properties of co(mpletely) positive matrices. *Linear Multilinear Algebra* 63(2): 384–396.

Shapiro A (2003) Monte Carlo sampling methods. Ruszczyński A, Shapiro A, eds. *Stochastic Programming*, Handbooks in Operations Research and Management Science, Vol. 10 (Elsevier, Amsterdam), 353–425.

Shapiro A, Kleywegt A (2002) Minimax analysis of stochastic problems. *Optim. Methods Software* 17(3):523–542.

Shapiro A, Dentcheva D, Ruszczyński A (2014) *Lectures on Stochastic Programming: Modeling and Theory* (SIAM, Philadelphia).

Sion M (1958) On general minimax theorems. *Pacific J. Math.* 8(1): 171–176.

Steinberg D (2005) Computation of matrix norms with applications to robust optimization. Unpublished master's thesis, Technion–Israel Institute of Technology, Haifia.

Thiele A, Terry T, Epelman M (2009) Robust linear optimization with recourse. Working paper, Lehigh University, Bethlehem, PA.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* 58(1):267–288.

Villani C (2008) *Optimal Transport: Old and New* (Springer-Verlag, Berlin).

Wang L, Gordon MD, Zhu J (2006) Regularized least absolute deviations regression and an efficient algorithm for parameter tuning. Clifton CW, Zhong N, Liu J, Wah BW, Wu X, eds. *IEEE Sixth Internat. Conf. Data Mining* (IEEE Computer Society, Los Alamitos, CA), 690–700.

Wiesemann W, Kuhn D, Sim M (2014) Distributionally robust convex optimization. *Oper. Res.* 62(6):1358–1376.

Wozabal D (2012) A framework for optimization under ambiguity. *Ann. Oper. Res.* 193(1):21–47.

Xu G, Burer S (2018) A copositive approach for two-stage adjustable robust optimization with uncertain right-hand sides. *Comput. Optim. Appl.* 70(1):33–59.

Zeng B, Zhao L (2013) Solving two-stage robust optimization problems using a column-and-constraint generation method. *Oper. Res. Lett.* 41(5):457–461.

Zhang D, Shen D, Alzheimer's Disease Neuroimaging Initiative (2012) Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage* 59(2):895–907.

Zhao C, Guan Y (2018) Data-driven risk-averse stochastic optimization with Wasserstein metric. *Oper. Res. Lett.* 46(2):262–267.

Grani A. Hanasusanto is an assistant professor of operations research and industrial engineering at the University of Texas at Austin. His research focuses on the design and analysis of tractable solution schemes for decision-making problems under uncertainty, with applications in operations management, energy systems, machine learning, and data analytics.

Daniel Kuhn is professor of operations research at the College of Management of Technology at EPFL, where he holds the Chair of Risk Analytics and Optimization. His current research interests are focused on the modeling of uncertainty, the development of efficient computational methods for the solution of stochastic and robust optimization problems, and the design of approximation schemes that ensure their computational tractability. This work is primarily application-driven, the main application areas being energy systems, operations management and engineering.