Data-Driven Chance Constrained Programs over Wasserstein Balls

Zhi Chen
College of Business, City University of Hong Kong, Kowloon Tong, Hong Kong,
zhi.chen@cityu.edu.hk

Daniel Kuhn
Risk Analytics and Optimization Chair, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland,
daniel.kuhn@epfl.ch

Wolfram Wiesemann
Imperial College Business School, Imperial College London, London, United Kingdom,
ww@imperial.ac.uk

We provide an exact deterministic reformulation for data-driven chance constrained programs over Wasserstein balls. For individual chance constraints as well as joint chance constraints with right-hand side uncertainty, our reformulation amounts to a mixed-integer conic program. In the special case of a Wasserstein ball with the 1-norm or the ∞-norm, the cone is the nonnegative orthant, and the chance constrained program can be reformulated as a mixed-integer linear program. Our reformulation compares favourably to several state-of-the-art data-driven optimization schemes in our numerical experiments.

Key words: Distributionally robust optimization; ambiguous chance constraints; Wasserstein distance.
History: May 9, 2022

1. Introduction

Distributionally robust optimization is a powerful modeling paradigm for optimization under uncertainty, where the distribution of the uncertain problem parameters is itself uncertain, and where the performance of a decision is assessed in view of the worst-case distribution from a prescribed ambiguity set. The earlier literature on distributionally robust optimization has focused on moment ambiguity sets which contain all distributions that obey certain (standard or generalized) moment conditions; see, *e.g.*, Delage and Ye (2010), Goh and Sim (2010) and Wiesemann et al. (2014). Pflug and Wozabal (2007) were the first to propose an ambiguity set of the form of a ball in the space of distributions with respect to the celebrated Wasserstein, Kanthorovich or optimal transport distance. The type-1 Wasserstein distance $d_{\mathrm{W}}(\mathbb{P}_1, \mathbb{P}_2)$ between two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ on $\mathbb{R}^K$, equipped with a general norm $\|\cdot\|$, is defined as the minimal transportation cost of moving

$\mathbb{P}_1$ to $\mathbb{P}_2$ under the premise that the cost of moving a Dirac point mass from ξ_1 to ξ_2 amounts to $\|\xi_1 - \xi_2\|$. Mathematically, this implies that

$$d_{\mathrm{W}}(\mathbb{P}_1, \mathbb{P}_2) = \inf_{\mathbb{P} \in \mathcal{P}(\mathbb{P}_1, \mathbb{P}_2)} \mathbb{E}_{\mathbb{P}}[\|\tilde{\xi}_1 - \tilde{\xi}_2\|],$$

where $\tilde{\xi}_1 \sim \mathbb{P}_1$, $\tilde{\xi}_2 \sim \mathbb{P}_2$, and $\mathcal{P}(\mathbb{P}_1, \mathbb{P}_2)$ represents the set of all distributions on $\mathbb{R}^K \times \mathbb{R}^K$ with marginals $\mathbb{P}_1$ and $\mathbb{P}_2$. The Wasserstein ambiguity set $\mathcal{F}(\theta)$ is then defined as a ball of radius $\theta \geq 0$ with respect to the Wasserstein distance, centered at a prescribed reference distribution $\hat{\mathbb{P}}$:

$$\mathcal{F}(\theta) = \{\mathbb{P} \in \mathcal{P}(\mathbb{R}^K) \mid d_{\mathrm{W}}(\mathbb{P}, \hat{\mathbb{P}}) \leq \theta\}. \tag{1}$$

One can think of the Wasserstein radius θ as a budget on the transportation cost. Indeed, any member distribution in $\mathcal{F}(\theta)$ can be obtained by rearranging the reference distribution $\hat{\mathbb{P}}$ at a transportation cost of at most θ. If only a finite training dataset $\{\hat{\xi}_i\}_{i \in [N]}$ is available, a natural choice for $\hat{\mathbb{P}}$ is the empirical distribution $\hat{\mathbb{P}} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\hat{\xi}_i}$, which represents the uniform distribution on the training samples. Throughout the paper, we will assume that $\hat{\mathbb{P}}$ is the empirical distribution.

While it has been recognized early on that Wasserstein ambiguity sets offer many conceptual advantages (*e.g.*, their member distributions do not need to be absolutely continuous with respect to $\hat{\mathbb{P}}$ and, if properly calibrated, they constitute confidence regions for the unknown true data-generating distribution), it was believed that they almost invariably lead to hard global optimization problems. Recently, Mohajerin Esfahani and Kuhn (2018) and Zhao and Guan (2018) discovered that many interesting distributionally robust optimization problems over Wasserstein ambiguity sets can actually be reformulated as tractable convex programs—provided that $\hat{\mathbb{P}}$ is discrete and that the problem's objective function satisfies certain convexity properties. These reformulations have subsequently been generalized to Polish spaces and non-discrete reference distributions by Blanchet and Murthy (2019) and Gao and Kleywegt (2016). Since then, distributionally robust optimization models over Wasserstein ambiguity sets have been proposed for many applications, including transportation (Carlsson et al. 2018) and machine learning (Blanchet et al. 2019, Gao et al. 2017, Shafieezadeh-Abadeh et al. 2019 and Sinha et al. 2017).

In this paper we study distributionally robust chance constrained programs of the form

$$\begin{array}{ll} \min\limits_{x \in X} & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{s.t.} & \mathbb{P}[\tilde{\xi} \in \mathcal{S}(\boldsymbol{x})] \geq 1 - \varepsilon \quad \forall \mathbb{P} \in \mathcal{F}(\theta), \end{array} \tag{2}$$

where the goal is to find a decision $\boldsymbol{x}$ from within a compact polyhedron $\mathcal{X} \subseteq \mathbb{R}^L$ that minimizes a linear cost function $\boldsymbol{c}^\top \boldsymbol{x}$ and ensures that the exogenous random vector $\tilde{\xi}$ falls within a decision-dependent safety set $\mathcal{S}(\boldsymbol{x}) \subseteq \mathbb{R}^K$ with high probability $1 - \varepsilon$ under every distribution $\mathbb{P} \in \mathcal{F}(\theta)$.

Since the reference distribution $\hat{\mathbb{P}}$ in (2) is the empirical distribution over the training dataset $\{\hat{\ }_i\}_{i2[N]}$, we refer to (2) as a *data-driven* chance constrained program.

To date, the literature on data-driven chance constraints has focused primarily on variants of problem (2) where the Wasserstein ambiguity set $\mathcal{F}(\theta)$ is replaced with an ambiguity set $\mathcal{G}(\theta)$ that contains all distributions close to the empirical distribution $\hat{\mathbb{P}}$ with respect to a ϕ-divergence (such as the Kullback-Leibler divergence or the χ^2-distance):

$$\mathcal{G}(\theta) = \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}^K) \;\middle|\; \mathbb{P} \ll \hat{\mathbb{P}},\ \int_{\mathbb{R}^K} \phi\left(\frac{\mathrm{d}\mathbb{P}(\ )}{\mathrm{d}\hat{\mathbb{P}}(\ )}\right) \mathrm{d}\hat{\mathbb{P}}(\ ) \leq \theta \right\},$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}$ is the divergence function. Hu and Hong (2013) show that a distributionally robust chance constrained program over a Kullback-Leibler ambiguity set reduces to a classical chance constrained progam over the reference distribution $\hat{\mathbb{P}}$ and an adjusted risk threshold $\varepsilon' < \varepsilon$. While this result holds for any reference distribution, ϕ-divergence ambiguity sets only contain distributions that are absolutely continuous with respect to $\hat{\mathbb{P}}$, that is, any distribution in $\mathcal{G}(\theta)$ only assigns positive probability to those measurable subsets $A \subseteq \mathbb{R}^K$ for which $\hat{\mathbb{P}}[\tilde{\ } \in A] > 0$. This is undesirable for problems with a large dimension K and/or few training data, where it is unlikely that every possible value of $\tilde{\ }$ has been observed in $\{\hat{\ }_i\}_{i2[N]}$. This shortcoming is addressed by Jiang and Guan (2016, 2018), who replace the reference distribution with a Kernel density estimator.

Despite their tremendous success and widespread adoption in recent years, the use of ϕ-divergences can lead to undesirable side effects in some applications: they compare distributions on a "scenario-by-scenario" basis and thus do not consider the possibility of noisy measurements (Gao and Kleywegt 2016), and they generically fail to be probability metrics as they typically violate symmetry as well as the triangle inequality. Moreover, as we show next, ϕ-divergence ambiguity sets may be overly optimistic when only few training samples are available.

Motivating Example. Consider the arguably simplest instance of the data-driven optimization problem (2), which estimates the worst-case value-at-risk $\sup_{\mathbb{P}2\mathcal{F}(\theta)} \mathbb{P}\text{-VaR}_\varepsilon(\tilde{\xi})$ of a scalar random variable $\tilde{\xi}$ at level ε from a limited set of i.i.d. training samples $\{\hat{\xi}_i\}_{i=1}^N$ of $\tilde{\xi}$ under the unknown data-generating distribution $\mathbb{P}_0$ that are summarized by the empirical distribution $\hat{\mathbb{P}} = \frac{1}{N}\sum_{i=1}^N \delta_{\hat{\xi}_i}$ at the centre of the Wasserstein ball $\mathcal{F}(\theta)$. To avoid technicalities, we assume that $\mathbb{P}_0$ is atomless. In addition, with $N_\varepsilon = \lfloor (1-\varepsilon)N \rfloor$ and $N^\varepsilon = \lceil (1-\varepsilon)N \rceil$ we define a distribution

$$\mathbb{P}^\varepsilon = \frac{1}{N}\sum_{i=1}^{N_\varepsilon} \delta_{\hat{\xi}_{(i)}} + \frac{(1-\varepsilon)N - N_\varepsilon}{N} \delta_{\hat{\xi}_{(N^\varepsilon)}} + \frac{N^\varepsilon - (1-\varepsilon)N}{N} \delta_{\hat{\xi}_{(N^\varepsilon)}+\theta/\varepsilon} + \frac{1}{N}\sum_{i=N^\varepsilon+1}^{N} \delta_{\hat{\xi}_{(i)}+\theta/\varepsilon}$$

to be used subsequently. Here, $\hat{\xi}_{(j)}$ denotes the j-th order statistic of the training samples $\{\hat{\xi}_i\}_{i=1}^N$.

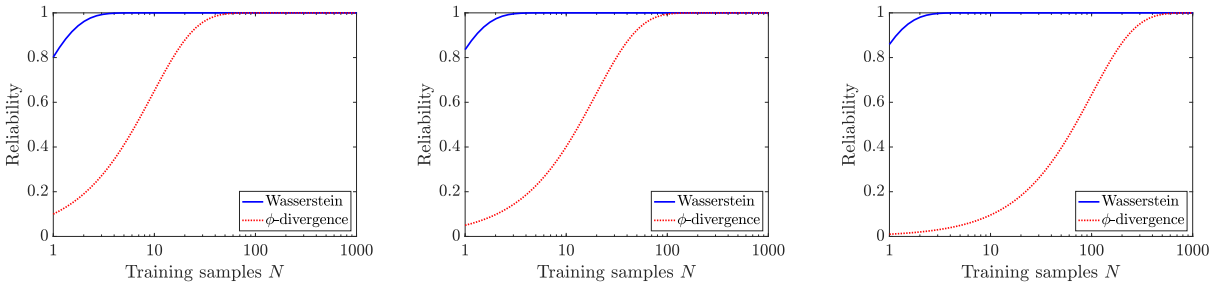

Figure 1 Reliability bounds for the Wasserstein (worst-case) and -divergence (best-case) ambiguity sets when approximating the VaR at level $\varepsilon = 0.1$ (left), $\varepsilon = 0.05$ (middle) and $\varepsilon = 0.01$ (right). We choose the radius $\theta = 1/\sqrt{N}$ for the Wassestein ball (see, *e.g.*, Mohajerin Esfahani and Kuhn 2018).

The *reliability* of the aforementioned worst-case value-at-risk, that is, the probability that it weakly exceeds the unknown true value-at-risk $\mathbb{P}_0\text{-VaR}_\varepsilon(\tilde{\xi})$, can be bounded from *below* by

$$\begin{aligned}
\mathbb{P}_0^N\left[\sup_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{P}\text{-VaR}_\varepsilon(\tilde{\xi}) \geq \mathbb{P}_0\text{-VaR}_\varepsilon(\tilde{\xi})\right] &\geq \mathbb{P}_0^N\left[\mathbb{P}^y\text{-VaR}_\varepsilon(\tilde{\xi}) \geq \mathbb{P}_0\text{-VaR}_\varepsilon(\tilde{\xi})\right] \\
&= \mathbb{P}_0^N\left[\hat{\mathbb{P}}\text{-VaR}_\varepsilon(\tilde{\xi}) \geq \mathbb{P}_0\text{-VaR}_\varepsilon(\tilde{\xi}) - \theta/\varepsilon\right] \\
&= 1 - \mathbb{P}_0^N\left[\hat{\mathbb{P}}\text{-VaR}_\varepsilon(\tilde{\xi}) < \mathbb{P}_0\text{-VaR}_\varepsilon(\tilde{\xi}) - \theta/\varepsilon\right] \\
&\geq 1 - \exp\left(-2N(1-\varepsilon-\mathbb{P}_0[\tilde{\xi} \leq \mathbb{P}_0\text{-VaR}_\varepsilon(\tilde{\xi}) - \theta/\varepsilon])^2\right),
\end{aligned}$$

where $\mathbb{P}_0^N$ is the N-fold product of $\mathbb{P}_0$ that generates $\{\hat{\xi}_i\}_{i=1}^N$. The first inequality holds since $\mathbb{P}^y$ is contained in $\mathcal{F}(\theta)$. The first equality holds since $\mathbb{P}^y\text{-VaR}_\varepsilon(\tilde{\xi}) = \hat{\mathbb{P}}\text{-VaR}_\varepsilon(\tilde{\xi}) + \theta/\varepsilon$ by construction of $\mathbb{P}^y$, and the last inequality is due to a standard concentration inequality for empirical quantiles (see, *e.g.*, Theorem 2.3.2 of Serfling 2009).

If we replace the Wasserstein ambiguity set $\mathcal{F}(\theta)$ with the ambiguity $\mathcal{G}(\theta)$ of any ϕ-divergence, on the other hand, then we can bound the reliability from *above* by

$$\begin{aligned}
\mathbb{P}_0^N\left[\sup_{\mathbb{P}\in\mathcal{G}(\theta)} \mathbb{P}\text{-VaR}_\varepsilon(\tilde{\xi}) \geq \mathbb{P}_0\text{-VaR}_\varepsilon(\tilde{\xi})\right] &= 1 - \mathbb{P}_0^N\left[\sup_{\mathbb{P}\in\mathcal{G}(\theta)} \mathbb{P}\text{-VaR}_\varepsilon(\tilde{\xi}) < \mathbb{P}_0\text{-VaR}_\varepsilon(\tilde{\xi})\right] \\
&\leq 1 - \mathbb{P}_0^N\left[\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_N < \mathbb{P}_0\text{-VaR}_\varepsilon(\tilde{\xi})\right] \\
&\leq 1 - (1-\varepsilon)^N.
\end{aligned}$$

Here, the first inequality holds since all distributions in $\mathcal{G}(\theta)$ share a common support with $\hat{\mathbb{P}}$, and the second inequality follows from the definition of $\mathbb{P}_0\text{-VaR}_\varepsilon(\tilde{\xi})$. We highlight that this probability bound holds for every radius θ of the ϕ-divergence ball $\mathcal{G}(\theta)$.

Figure 1 compares the *worst-case* reliability offered by the Wasserstein ambiguity set with the *best-case* reliability of the ϕ-divergence ambiguity set for a uniform distribution over the interval $[0, 1]$. We observe that in low-sample regimes, ϕ-divergence ambiguity sets may underestimate the true VaR with high probability. ♣

To our best knowledge, the paper of Xie and Ahmed (2020) is the only previous work on data-driven chance constraints over Wasserstein ambiguity sets. The authors study the special class of covering problems, where the feasible region $\mathcal{X}$ satisfies $\eta\mathcal{X} \subseteq \mathcal{X}$ for every $\eta \geq 1$. This problem class encompasses, among others, portfolio optimization problems without budgetary restrictions and lot-sizing problems in the absence of setup costs. The authors prove that the resulting individual chance constrained program is NP-hard. They also demonstrate that two popular approximation schemes, the CVaR approximation as well as the scenario approximation, can perform arbitrarily poorly for classical individual chance constraints, that is, when the Wasserstein radius is $\theta = 0$. Based on this insight, the authors propose a bicriteria approximation scheme for covering problems with classical as well as distributionally robust individual chance constraints over moment and Wasserstein ambiguity sets. This bicriteria approximation scheme determines solutions that trade off a higher risk threshold $\varepsilon' > \varepsilon$ in the chance constraint with a smaller optimality gap $\varepsilon'/(\varepsilon' - \varepsilon)$. This is achieved by solving a tractable convex relaxation of the chance constrained problem (using, *e.g.*, a Markovian or Bernstein generator) and subsequently scaling the solution to this relaxation so that it becomes feasible for the chance constraint with the higher risk threshold ε'. By design, the performance guarantee of the bicriteria approximation scheme becomes weaker (and eventually trivial) as the selected risk threshold ε' approaches the risk threshold ε of the original problem formulation.

In this paper, we study distributionally robust chance constrained programs over the Wasserstein ambiguity set (1). We derive deterministic reformulations for individual chance constrained programs, where $\mathcal{S}(\boldsymbol{x}) = \{\ \in \mathbb{R}^K \mid \boldsymbol{a}(\ )^\top \boldsymbol{x} < b(\ )\}$ for affine functions $\boldsymbol{a}(\cdot) : \mathbb{R}^K \to \mathbb{R}^L$ and $b(\cdot) : \mathbb{R}^K \to \mathbb{R}$, as well as for joint chance constrained programs with right-hand side uncertainty, where $\mathcal{S}(\boldsymbol{x}) = \{\ \in \mathbb{R}^K \mid \boldsymbol{A}\boldsymbol{x} < \boldsymbol{b}(\ )\}$ for $\boldsymbol{A} \in \mathbb{R}^{M\ \ L}$ and an affine function $\boldsymbol{b} : \mathbb{R}^K \to \mathbb{R}^M$. Our reformulations are mixed-integer conic programs that reduce to mixed-integer linear programs when the norm $\|\cdot\|$ on $\mathbb{R}^K$ is the 1-norm or the ∞-norm.

While preparing this paper for publication, we became aware of the independent work by Xie (2019), which derives similar reformulations for distributionally individual and joint chance constraints over Wasserstein ambiguity sets. In contrast to our work, however, Xie (2019) assumes that each safety condition $\boldsymbol{a}_m^\top \boldsymbol{x} < b_m(\ )$, $m \in [M]$, in the joint chance constraint depends on a subvector of , and that these subvectors are pairwise disjoint for different safety conditions. In other words, different safety conditions of the joint chance constraints studied by Xie (2019) must depend on different random variables. Furthermore, the reformulations of Xie (2019) are derived via duality theory, whereas our reformulations directly leverage the structural insights into the worst-case distributions. This enables us to keep our reformulations largely independent of the selected ground metric for the Wasserstein ball, which opens up possibilities to incorporate other cost functions

in our definition of the Wasserstein distance. Since the initial submission of this paper, our exact reformulation for data-driven chance constrained program over Wasserstein balls has been further studied and tightened; see, for instance, Ho-Nguyen et al. (2020, 2021), Shen and Jiang (2021) and Zhang and Dong (2021). Along with these theoretical extensions, our reformulation has also been applied in several domains, including risk sharing in finance (Chen and Xie 2021), network design for humanitarian operations (Jiang et al. 2021) and optimal power flows in energy systems (Arrigo et al. 2022).

Notation. We use boldface uppercase and lowercase letters to denote matrices and vectors, respectively. Special vectors of appropriate dimensions include $\mathbf{0}$ and $\mathbf{e}$, which respectively correspond to the zero vector and the vector of all ones. We denote by $\|\cdot\|_*$ the dual norm of a general norm $\|\cdot\|$. We use the shorthand $[N] = \{1, 2, \ldots, N\}$ to represent the set of all integers up to N. Given a (possibly fractional) real number $\ell \in [0, N]$, we define the partial sum of the ℓ first values in $\{k_i\}_{i \in [N]}$ as $\sum_{i=1}^{\ell} k_i = \sum_{i=1}^{\lfloor \ell \rfloor} k_i + (\ell - \lfloor \ell \rfloor) k_{\lfloor \ell \rfloor + 1}$. Random vectors are denoted by tilde signs (*e.g.*, $\tilde{\ }$), while their realizations are denoted by the same symbols without tildes (*e.g.*,). Given a random vector $\tilde{\ }$ governed by a distribution $\mathbb{P}$, a measurable loss function $\ell(\ )$ and a risk threshold $\varepsilon \in (0, 1)$, the value-at-risk (VaR) of $\ell(\ )$ at level ε is defined as $\mathbb{P}\text{-VaR}_\varepsilon(\ell(\ )) = \inf\{\gamma \in \mathbb{R} \mid \mathbb{P}[\gamma \leq \ell(\tilde{\ })] \leq \varepsilon\}$.

2. Exact Reformulation of Data-Driven Chance Constraints

Section 2.1 reviews a previously established result on the quantification of uncertainty over Wasserstein balls. We use this result to derive an exact reformulation of generic data-driven chance constrained programs in Section 2.2. We finally specialize this generic reformulation to the subclasses of data-driven individual chance constrained programs as well as data-driven joint chance constrained programs with right-hand side uncertainty in Sections 2.3 and 2.4, respectively.

2.1. Uncertainty Quantification over Wasserstein Balls

Consider an open safety set $\mathcal{S} \subseteq \mathbb{R}^K$, and denote by $\bar{\mathcal{S}} = \mathbb{R}^K \setminus \mathcal{S}$ its closed complement. The uncertainty quantification problem

$$\sup_{\mathbb{P} \in \mathcal{F}(\theta)} \mathbb{P}[\tilde{\ } \notin \mathcal{S}] \tag{3}$$

computes the worst (largest) probability of the system under consideration being unsafe, which is the case whenever the random vector $\tilde{\ }$ attains a value in the unsafe set $\bar{\mathcal{S}}$. Throughout the rest of the paper, we exclude trivial special cases and assume that $\theta > 0$ and $\varepsilon \in (0, 1)$.

To solve the uncertainty quantification problem (3), denote by $\mathbf{dist}(\hat{\ }_i, \bar{\mathcal{S}})$ the distance of the i^{th} data point $\hat{\ }_i \in \mathbb{R}^K$ of the empirical distribution $\hat{\mathbb{P}}$ to the unsafe set $\bar{\mathcal{S}}$. This distance is based on a norm $\|\cdot\|$, which we keep generic at this stage. Without loss of generality, we assume that the

data points $\{\hat{\xi}_i\}_{i\in[N]}$ are ordered in increasing distance to $\bar{\mathcal{S}}$, that is, $\mathbf{dist}(\hat{\xi}_i, \bar{\mathcal{S}}) \leq \mathbf{dist}(\hat{\xi}_j, \bar{\mathcal{S}})$ for all $1 \leq i \leq j \leq N$. We also assume that $\mathbf{dist}(\hat{\xi}_i, \bar{\mathcal{S}}) = 0$ (that is, the data point $\hat{\xi}_i$ is unsafe) if and only if $i \in [I]$, where $I = 0$ if $\mathbf{dist}(\hat{\xi}_i, \bar{\mathcal{S}}) > 0$ for all $i \in [N]$. Finally, we denote by $\xi_i^\star \in \bar{\mathcal{S}}$ an unsafe point that is closest to the data point $\hat{\xi}_i$, $i \in [N]$, in terms of the distance $\mathbf{dist}(\hat{\xi}_i, \bar{\mathcal{S}})$.

Blanchet and Murthy (2019) as well as Gao and Kleywegt (2016) have characterized the solution to the uncertainty quantification problem (3) in closed form. To keep our paper self-contained, we reproduce their findings without proof in Theorem 1 below.

THEOREM 1. *Let* $j^\star = \max\{j \in [N] \cup \{0\} \mid \sum_{i=1}^{j} \mathbf{dist}(\hat{\xi}_i, \bar{\mathcal{S}}) \leq \theta N\}$. *The uncertainty quantification problem* (3) *is solved by a worst-case distribution* $\mathbb{P}^\star \in \mathcal{F}(\theta)$ *that is characterized as follows:*

(i) If $j^\star = N$, *then* $\sup_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{P}[\tilde{\xi} \notin \mathcal{S}] = \mathbb{P}^\star[\tilde{\xi} \notin \mathcal{S}] = 1$ *for*

$$\mathbb{P}^\star = \frac{1}{N}\sum_{i=1}^{I} \delta_{\hat{\xi}_i} + \frac{1}{N}\sum_{i=I+1}^{N} \delta_{\xi_i^\star}.$$

(ii) If $j^\star < N$, *then* $\sup_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{P}[\tilde{\xi} \notin \mathcal{S}] = \mathbb{P}^\star[\tilde{\xi} \notin \mathcal{S}] = (j^\star + p^\star)/N$ *for*

$$\mathbb{P}^\star = \frac{1}{N}\sum_{i=1}^{I} \delta_{\hat{\xi}_i} + \frac{1}{N}\sum_{i=I+1}^{j^\star} \delta_{\xi_i^\star} + \frac{p^\star}{N}\delta_{\xi_{j^\star+1}^\star} + \frac{1-p^\star}{N}\delta_{\hat{\xi}_{j^\star+1}} + \frac{1}{N}\sum_{i=j^\star+2}^{N} \delta_{\hat{\xi}_i},$$

where $p^\star = (\theta N - \sum_{i=1}^{j^\star} \mathbf{dist}(\hat{\xi}_i, \bar{\mathcal{S}}))/\mathbf{dist}(\hat{\xi}_{j^\star+1}, \bar{\mathcal{S}})$.

Intuitively speaking, the worst-case distribution $\mathbb{P}^\star$ in Theorem 1 transports the training dataset $\{\hat{\xi}_i\}_{i\in[N]}$ to the unsafe set $\bar{\mathcal{S}}$ in a greedy fashion, see Figure 2. The data points $\hat{\xi}_1, \ldots, \hat{\xi}_I$ are already unsafe and hence do not need to be transported. The subsequent data points $\hat{\xi}_{I+1}, \ldots, \hat{\xi}_{j^\star+1}$ are closest to the unsafe set and are thus transported from $\mathcal{S}$ to $\bar{\mathcal{S}}$. Due to the limited transportation budget θ, the data point $\hat{\xi}_{j^\star+1}$ is only partially transported. The safe samples $\hat{\xi}_{j^\star+2}, \ldots \hat{\xi}_N$, finally, are too far away from the unsafe set $\bar{\mathcal{S}}$ and are thus left unchanged. Note that the distribution characterized in Theorem 1 may not be the only distribution that solves problem (3).

2.2. Reformulation of Generic Chance Constraints

We now develop deterministic reformulations for the distributionally robust chance constrained program (2). To this end, we focus on the ambiguous chance constraint

$$\sup_{\mathbb{P}\in\mathcal{F}(\theta)} \mathbb{P}[\tilde{\xi} \notin \mathcal{S}(x)] \leq \varepsilon. \tag{4}$$

For any fixed decision $x \in \mathcal{X}$, we let $\mathcal{S}(x)$ be an arbitrary open safety set, and we denote by $\bar{\mathcal{S}}(x)$ its closed complement, which comprises all unsafe scenarios. Every fixed training dataset $\{\hat{\xi}_i\}_{i\in[N]}$

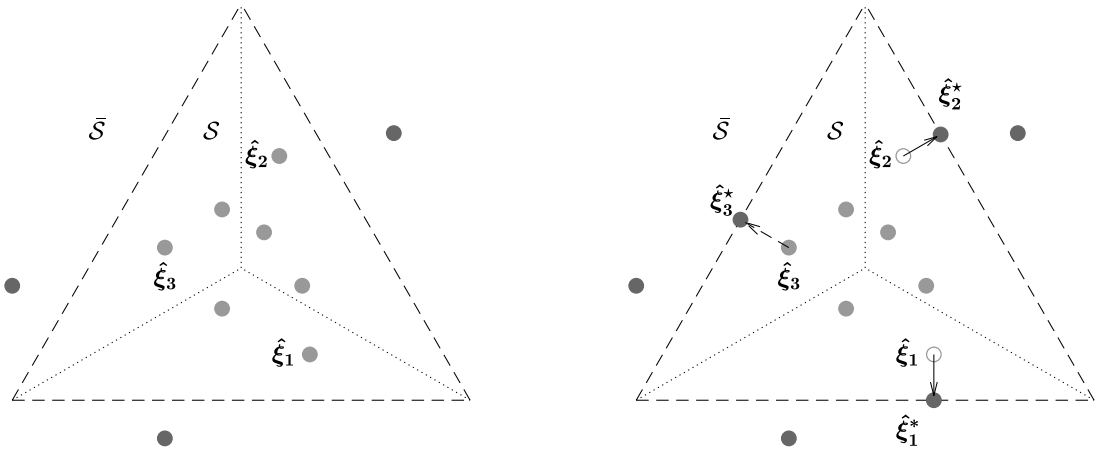

Figure 2 Empirical and worst-case distributions. The left graph visualizes the empirical distribution $\hat{\mathbb{P}}$, whose light grey (dark grey) data points are contained in (outside of) the safety set $\mathcal{S}$ shown as an equilateral triangle (dashed lines). The right graph shows the corresponding worst-case distribution $\mathbb{P}^\star$, which moves the data points $\hat{\xi}_1$ and $\hat{\xi}_2$ entirely as well as the data point $\hat{\xi}_3$ partially to the unsafe set $\bar{\mathcal{S}}$. Each transported data point is projected onto the boundary of the closest halfspace defining the safety set $\mathcal{S}$.

then induces a (decision-dependent) permutation $\pi(\boldsymbol{x})$ of $[N]$ that orders the training samples in increasing distance to the unsafe set, that is,

$$\mathbf{dist}(\hat{\xi}_{\pi_1(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) \leq \mathbf{dist}(\hat{\xi}_{\pi_2(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) \leq \cdots \leq \mathbf{dist}(\hat{\xi}_{\pi_N(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})).$$

We first show that a fixed decision $\boldsymbol{x}$ satisfies the ambiguous chance constraint (4) over the Wasserstein ambiguity set (1) if and only if the partial sum of the εN smallest transportation distances to the unsafe set multiplied by the mass $1/N$ of a training sample exceeds θ.

THEOREM 2. *For any fixed decision $\boldsymbol{x} \in \mathcal{X}$, the ambiguous chance constraint* (4) *over the Wasserstein ambiguity set* (1) *is equivalent to the deterministic inequality*

$$\frac{1}{N} \sum_{i=1}^{\varepsilon N} \mathbf{dist}(\hat{\xi}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) \geq \theta. \tag{5}$$

The left-hand side of (5) can be interpreted as the minimum cost of moving a fraction ε of the training samples to the unsafe set. If this cost exceeds the prescribed transportation budget θ, then no distribution in the Wasserstein ambiguity set can assign the unsafe set a probability of more than ε, which means that the distributionally robust chance constraint (4) is satisfied.

Proof of Theorem 2. From Theorem 1 we know that the worst-case distribution $\mathbb{P}^\star$ is an optimal solution (not necessarily unique) to the maximization problem embedded in the left-hand side of the ambiguous chance constraint (4). We thus conclude that the constraint (4) is satisfied if and only if $\mathbb{P}^\star[\tilde{\xi} \notin \mathcal{S}(\boldsymbol{x})] \leq \varepsilon$ for $\mathbb{P}^\star$ defined in the statement of that theorem.

In case (i) of Theorem 1, the ambiguous chance constraint (4) is violated since $\mathbb{P}^\star[\tilde{\boldsymbol{x}} \notin \mathcal{S}(\boldsymbol{x})] = 1$ while $\varepsilon < 1$ by assumption. At the same time, since $j^\star = N$, we have $\frac{1}{N}\sum_{i=1}^{N} \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) \leq \theta$. If this inequality is strict, then (5) is violated as desired since $\frac{1}{N}\sum_{i=1}^{\varepsilon N} \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) \leq \frac{1}{N}\sum_{i=1}^{N} \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x}))$. If the inequality is satisfied as an equality, on the other hand, we know that $\mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_N(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) > 0$ since $\theta > 0$ by assumption and $\mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) \leq \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_j(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x}))$ for all $i \leq j$ by construction of the re-ordering $(\boldsymbol{x})$. Thus, since $\varepsilon < 1$ by assumption, we have $\frac{1}{N}\sum_{i=1}^{\varepsilon N} \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) < \frac{1}{N}\sum_{i=1}^{N} \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) = \theta$ and equation (5) is violated as desired.

In case (ii) of Theorem 1, we have $\mathbb{P}^\star[\tilde{\boldsymbol{x}} \notin \mathcal{S}(\boldsymbol{x})] = (j^\star + p^\star)/N$ with $j^\star = \max\{j \in [N-1] \cup \{0\} \mid \sum_{i=1}^{j} \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) \leq \theta N\}$ as well as $p^\star = (\theta N - \sum_{i=1}^{j^\star} \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})))/\mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_{j^\star+1}(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x}))$. We claim that $j^\star + p^\star$ is the optimal value of the bivariate mixed-integer optimization problem

$$\begin{array}{ll} \max\limits_{j,p} & j + p \\ \text{s.t.} & \sum\limits_{i=1}^{j} \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) + p \cdot \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_{j+1}(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) \leq \theta N \\ & j \in [N-1] \cup \{0\},\ 0 \leq p < 1. \end{array} \tag{6}$$

Indeed, the solution $(j, p) = (j^\star, p^\star)$ is feasible in (6) by definition of $j^\star$ and $p^\star$. Moreover, we have $j + p < j^\star + p^\star$ for any other feasible solution (j, p) that satisfies $j = j^\star$ and $p \neq p^\star$. Assume now that the optimal solution (j, p) to (6) would satisfy $j > j^\star$. Any such solution would violate the first constraint since $\sum_{i=1}^{j} \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) > \theta N$ by definition of $j^\star$ while $p \geq 0$. Similarly, any solution (j, p) with $j < j^\star$ cannot be optimal in (6) since $j \leq j^\star - 1$ while $p < p^\star + 1$.

We can re-express problem (6) as the univariate discrete optimization problem

$$\max\left\{ j \in [0, N] \;\middle|\; \sum_{i=1}^{\lfloor j \rfloor} \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) \;+\; (j - \lfloor j \rfloor) \cdot \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_{\lfloor j \rfloor + 1}(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) \leq \theta N \right\}.$$

Using our definition of partial sums, we observe that this problem is equivalent to

$$\max\left\{ j \in [0, N] \;\middle|\; \sum_{i=1}^{j} \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) \leq \theta N \right\}.$$

By construction, the mapping $\vartheta(j) = \sum_{i=1}^{j} \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x}))$, $j \in [0, N]$, is continuous and monotonically nondecreasing. It therefore affords the right inverse $\vartheta^{-1}(t) = \max\{j \in [0, N] \mid \vartheta(j) \leq t\}$ that satisfies $\vartheta \circ \vartheta^{-1}(t) = t$ for all $t \in [0, \vartheta(N)]$. Figure 3 visualizes the relationship between ϑ and ϑ^{-1}. We thus conclude that the ambiguous chance constraint (4) is satisfied if and only if

$$\begin{array}{rcl} \max\left\{ j \in [0, N] \;\middle|\; \sum\limits_{i=1}^{j} \mathbf{dist}(\hat{\boldsymbol{x}}_{\pi_i(\boldsymbol{x})}, \bar{\mathcal{S}}(\boldsymbol{x})) \leq \theta N \right\} \leq \varepsilon N & \Longleftrightarrow & \max\{j \in [0, N] \mid \vartheta(j) \leq \theta N\} \leq \varepsilon N \\ & \Longleftrightarrow & \vartheta^{-1}(\theta N) \leq \varepsilon N \\ & \Longleftrightarrow & \theta N \leq \vartheta(\varepsilon N), \end{array}$$

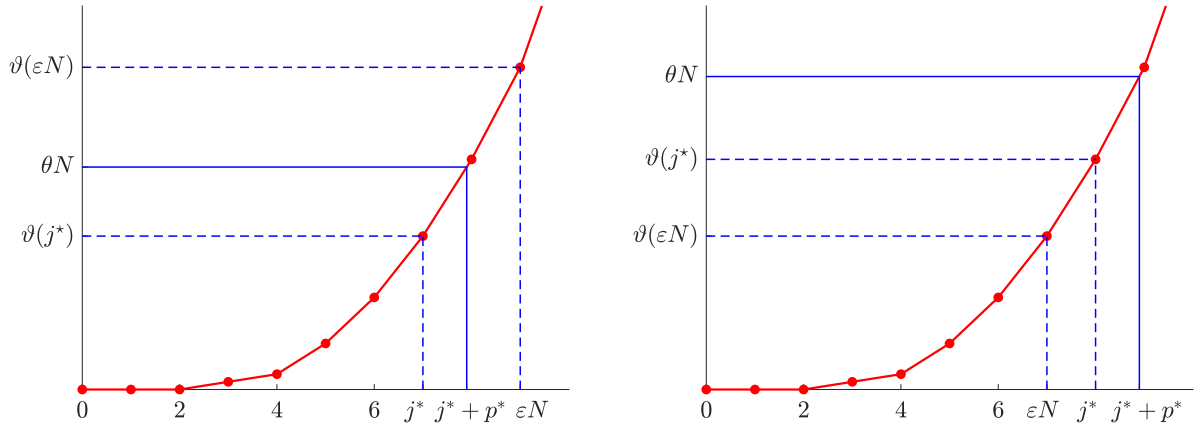

Figure 3 Relationship between # and # ^{1}. The left graph shows a feasible solution $\boldsymbol{x}$ satisfying the ambiguous chance constraint (4); in this case, we have #("N) $\geq$ N. The infeasible solution $\boldsymbol{x}^0$ in the right graph, on the other hand, violates the ambiguous chance constraint (4), and we have #("N) $<$ N.

where the last equivalence follows from $\vartheta \circ \vartheta^{-1}(\theta N) = \theta N$, which holds because $\theta N \leq \vartheta(N)$ for $j^\star < N$, as well as the fact that ϑ is monotonically nondecreasing. By definition, the right-hand side of the last equivalence holds if and only if (5) in the statement of the theorem is satisfied.

Remark 1. We emphasize that the inequality (5) fails to be equivalent to the ambiguous chance constraint (4) when $\theta = 0$, in which case the Wasserstein ball collapses to the singleton set $\mathcal{F}(0) = \{\hat{\mathbb{P}}\}$. To see this, suppose that $\hat{}_{\pi_i(x)} \in \bar{\mathcal{S}}(\boldsymbol{x})$ for all $i = 1, \ldots, I$ and $\hat{}_{\pi_i(x)} \in \mathcal{S}(\boldsymbol{x})$ for all $i = I+1, \ldots, N$, where $I \geq 1$. If $\varepsilon < I/N$, then the chance constraint (4) is violated because

$$\hat{\mathbb{P}}[\tilde{} \notin \mathcal{S}(\boldsymbol{x})] = \frac{I}{N} > \varepsilon,$$

while the inequality (5) holds trivially because $\sum_{i=1}^{\varepsilon N} \mathbf{dist}(\hat{}_{\pi_i(x)}, \bar{\mathcal{S}}(\boldsymbol{x})) \geq 0$.

Theorem 2 establishes that a decision $\boldsymbol{x} \in \mathcal{X}$ satisfies the ambiguous chance constraint (4) if and only if the sum of the εN smallest distances of the training samples to the unsafe set $\bar{\mathcal{S}}(\boldsymbol{x})$ weakly exceeds θN. This result is of computational interest because the sum of the εN smallest out of N real numbers is concave in those real numbers (while being convex in ε). This reveals that the constraint (5) is convex in the decision-dependent distances $\{\mathbf{dist}(\hat{}_i, \bar{\mathcal{S}}(\boldsymbol{x}))\}_{i2[N]}$. In the remainder we develop an efficient reformulation of this convex constraint that does not require an enumeration of all possible sums of εN different distances between the training samples and the unsafe set. This reformulation is based on the following auxiliary lemma.

LEMMA 1. *For any $\varepsilon \in (0,1)$, the sum of the εN smallest out of N real numbers $k_1, \ldots, k_N$ coincides with the optimal value of the linear program*

$$\begin{array}{ll} \max\limits_{s,t} & \varepsilon N t - \boldsymbol{e}^\top \boldsymbol{s} \\ \text{s.t.} & k_i \geq t - s_i \quad \forall i \in [N] \\ & \boldsymbol{s} \geq \boldsymbol{0}. \end{array}$$

Proof of Lemma 1. By definition, the sum of the εN smallest elements of the set $\{k_1, \ldots, k_N\}$ corresponds to the optimal value of the (manifestly feasible) linear program

$$\begin{array}{ll} \min\limits_{v} & \sum\limits_{i2[N]} k_i v_i \\ \text{s.t.} & \boldsymbol{0} \leq \boldsymbol{v} \leq \boldsymbol{e},\ \boldsymbol{e}^\top \boldsymbol{v} = \varepsilon N. \end{array}$$

The claim now follows from strong linear programming duality.

Armed with Theorem 2 and Lemma 1, we are now ready to reformulate the chance constrained program (2) as a deterministic optimization problem.

THEOREM 3. *The chance constrained program* (2) *is equivalent to*

$$\begin{array}{ll} \min\limits_{s,t,x} & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{s.t.} & \varepsilon N t - \boldsymbol{e}^\top \boldsymbol{s} \geq \theta N \\ & \mathbf{dist}(\hat{\ }_i, \bar{\mathcal{S}}(\boldsymbol{x})) \geq t - s_i \quad \forall i \in [N] \\ & \boldsymbol{s} \geq \boldsymbol{0},\ \boldsymbol{x} \in \mathcal{X}. \end{array} \tag{7}$$

Proof of Theorem 3. The claim follows immediately by using Theorem 2 to reformulate the chance constraint (4) as the inequality (5), using Lemma 1 to express the left-hand side of (5) as a linear maximization problem and substituting the resulting constraint back into (2).

We emphasize that the reformulation offered by Theorem 3 is independent of the selected ground metric $\mathbf{dist}(\cdot,\cdot)$. In the remainder, we assume that the ground metric is based on a norm $\|\cdot\|$.

2.3. Reformulation of Individual Chance Constraints

Assume now that problem (2) accommodates an individual chance constraint defined through the safety set $\mathcal{S}(\boldsymbol{x}) = \{\ \in \mathbb{R}^K \mid (\boldsymbol{A}\ + \boldsymbol{a})^\top \boldsymbol{x} < \boldsymbol{b}^\top\ + b\}$. Individual chance constrained programs have been studied, among others, in network design (Wang 2007), vehicle routing (Gounaris et al. 2013, Ghosal and Wiesemann 2020) and portfolio optimization (Rujeerapaiboon et al. 2016, Dert and Oldenkamp 2000). By Lemma 2 in the appendix, we have

$$\mathbf{dist}(\hat{\ }_i, \bar{\mathcal{S}}(\boldsymbol{x})) = \frac{((\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x})^\top \hat{\ }_i + b - \boldsymbol{a}^\top \boldsymbol{x})^+}{\|\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x}\|} \quad \forall i \in [N],$$

where we adopt the convention that $0/0 = 0$, and thus Theorem 3 allows us to reformulate problem (7) as the deterministic optimization problem

$$\begin{array}{ll} \min\limits_{s,t,x} & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{s.t.} & \varepsilon N t - \boldsymbol{e}^\top \boldsymbol{s} \geq \theta N \\ & \dfrac{((\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x})^\top \hat{}_i + b - \boldsymbol{a}^\top \boldsymbol{x})^+}{\|\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x}\|} \geq t - s_i \quad \forall i \in [N] \\ & \boldsymbol{s} \geq \boldsymbol{0}, \ \boldsymbol{x} \in \mathcal{X}. \end{array} \tag{8}$$

Unfortunately, problem (8) fails to be convex as its constraints involve fractions of convex functions. Below we show, however, that problem (8) can be reformulated as a mixed integer conic program.

Proposition 1. *Assume that $\boldsymbol{A}^\top \boldsymbol{x} \neq \boldsymbol{b}$ for all $\boldsymbol{x} \in \mathcal{X}$. For the safety set $\mathcal{S}(\boldsymbol{x}) = \{ \ \in \mathbb{R}^K \mid (\boldsymbol{A} \ + \boldsymbol{a})^\top \boldsymbol{x} < \boldsymbol{b}^\top \ + b\}$, problem (2) is equivalent to the mixed integer conic program*

$$\begin{array}{lll} Z^\star_{\mathrm{ICC}} = \min\limits_{q,s,t,x} & \boldsymbol{c}^\top \boldsymbol{x} & \\ \text{s.t.} & \varepsilon N t - \boldsymbol{e}^\top \boldsymbol{s} \geq \theta N \|\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x}\| & \\ & (\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x})^\top \hat{}_i + b - \boldsymbol{a}^\top \boldsymbol{x} + \mathrm{M} q_i \geq t - s_i \quad & \forall i \in [N] \\ & \mathrm{M}(1 - q_i) \geq t - s_i & \forall i \in [N] \\ & \boldsymbol{q} \in \{0,1\}^N, \ \boldsymbol{s} \geq \boldsymbol{0}, \ \boldsymbol{x} \in \mathcal{X}, & \end{array} \tag{9}$$

where M *is a suitably large (but finite) positive constant.*

Proof of Proposition 1. We already know that the chance constrained program (2) is equivalent to the non-convex optimization problem (8). A complicating feature of this problem is the appearance of the maximum operator in the second constraint group, which evaluates the positive part of $(\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x})^\top \hat{}_i + b - \boldsymbol{a}^\top \boldsymbol{x}$. To eliminate this maximum operator, for each $i \in [N]$ we introduce a binary variable $q_i \in \{0,1\}$, and we re-express the i^{th} member of the second constraint group via the two auxiliary constraints

$$\frac{(\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x})^\top \hat{}_i + b - \boldsymbol{a}^\top \boldsymbol{x}}{\|\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x}\|} + \mathrm{M} q_i \geq t - s_i \quad \text{and} \quad \mathrm{M}(1 - q_i) \geq t - s_i. \tag{10}$$

Note that at optimality we have $q_i = 1$ if $(\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x})^\top \hat{}_i + b - \boldsymbol{a}^\top \boldsymbol{x}$ is negative and $q_i = 0$ otherwise. Intuitively, q_i thus activates the less restrictive one of the two auxiliary constraints in (10). Next, we apply the variable substitutions $t \leftarrow t/\|\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x}\|$ and $\boldsymbol{s} \leftarrow \boldsymbol{s}/\|\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x}\|$, which is admissible because $\boldsymbol{A}^\top \boldsymbol{x} \neq \boldsymbol{b}$ for all $\boldsymbol{x} \in \mathcal{X}$. This change of variables yields the postulated reformulation (9).

To see that a finite value of M is sufficient for our reformulation to be exact, we show that the expression $((\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x})^\top \hat{}_i + b - \boldsymbol{a}^\top \boldsymbol{x})/\|\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x}\|$ as well as the values of t and s_i, $i \in [N]$, in (10) can all be bounded without affecting the optimal value of problem (9). This is clear for the fraction

as $\mathcal{X}$ is compact and the denominator is non-zero for all $\boldsymbol{x} \in \mathcal{X}$. Moreover, t is nonnegative as otherwise the first constraint in (9) would be violated. For any fixed values of $\boldsymbol{x}$ and t, an optimal value of s_i, $i \in [N]$, is given by $s_i^\star(\boldsymbol{x}, t) = (t - ((\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x})^\top \hat{}_i + b - \boldsymbol{a}^\top \boldsymbol{x})/\|\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x}\|_\star)^+$. Since $\mathcal{X}$ is bounded, it thus remains to show that t can be bounded from above. Indeed, for sufficiently large (but finite) t, the slope of $\varepsilon N t - \boldsymbol{e}^\top \boldsymbol{s}^\star(\boldsymbol{x}, t)$ on the left-hand side of the first constraint in (9) is $-(1-\varepsilon)N$. Since $\varepsilon < 1$, we thus conclude that this constraint is violated for large values of t.

Remark 2. The condition that $\boldsymbol{A}^\top \boldsymbol{x} \neq \boldsymbol{b}$ for all $\boldsymbol{x} \in \mathcal{X}$ does not restrict the generality of our formulation. Indeed, if an optimal solution $(\boldsymbol{q}^\star, \boldsymbol{s}^\star, t^\star, \boldsymbol{x}^\star)$ to problem (9) satisfies $\boldsymbol{A}^\top \boldsymbol{x}^\star \neq \boldsymbol{b}$, then $\boldsymbol{x}^\star$ solves problem (2) since our argument in the proof of Proposition 1 applies to $\boldsymbol{x}^\star$ even if $\boldsymbol{A}^\top \boldsymbol{x} = \boldsymbol{b}$ for some $\boldsymbol{x} \in \mathcal{X}$. Assume now that an optimal solution $(\boldsymbol{q}^\star, \boldsymbol{s}^\star, t^\star, \boldsymbol{x}^\star)$ to problem (9) satisfies $\boldsymbol{A}^\top \boldsymbol{x}^\star = \boldsymbol{b}$. In that case, the ambiguous chance constraint in problem (2) requires that $\boldsymbol{a}^\top \boldsymbol{x}^\star < b$. If that is the case for $\boldsymbol{x}^\star$, it is optimal in problem (2). If, finally, an optimal solution $(\boldsymbol{q}^\star, \boldsymbol{s}^\star, t^\star, \boldsymbol{x}^\star)$ to problem (9) satisfies $\boldsymbol{A}^\top \boldsymbol{x}^\star = \boldsymbol{b}$ and $\boldsymbol{a}^\top \boldsymbol{x}^\star \geq b$, then one would ideally like to solve a variant of problem (9) that includes the additional constraint

$$\boldsymbol{A}^\top \boldsymbol{x} \neq \boldsymbol{b} \quad \text{or} \quad \boldsymbol{a}^\top \boldsymbol{x} < b. \tag{11}$$

This variant of problem (9) can be solved by solving $2K+1$ versions of problem (9), where each version includes exactly one of the constraints $[\boldsymbol{A}^\top \boldsymbol{x}]_k > [\boldsymbol{b}]_k$, $[\boldsymbol{A}^\top \boldsymbol{x}]_k < [\boldsymbol{b}]_k$, $k \in [K]$, or $\boldsymbol{a}^\top \boldsymbol{x} < b$. One readily verifies that the solution that attains the least objective value amongst these $2K+1$ versions of problem (9) is an optimal solution to problem (9) with the added constraint (11).

Remark 3. The mixed-integer conic program (9) simplifies to a mixed-integer linear program whenever $\|\cdot\|$ represents the 1-norm or the ∞-norm, and it can be reformulated as a mixed-integer second-order cone program whenever $\|\cdot\|$ represents a p-norm for some $p \in \mathbb{Q}$, $p > 1$, see Section 2.3.1 in Ben-Tal and Nemirovski (2001).

Remark 4. The deterministic reformulation (9) is remarkably parsimonious. For an L-dimensional feasible region $\mathcal{X} \subseteq \mathbb{R}^L$ and an empirical distribution $\hat{\mathbb{P}}$ with N data points, our reformulation (9) has N binary variables, $L+N+1$ continuous decisions as well as $2N+1$ constraints (excluding those that describe $\mathcal{X}$). In comparison, a classical chance constrained formulation, which is tantamount to setting the Wasserstein radius to $\theta = 0$ in problem (2), has N binary variables, L continuous decisions as well as $N+1$ constraints. Thus, adding distributional robustness only requires an additional $N+1$ continuous decisions as well as N further constraints.

Remark 5. The deterministic reformulation (9) requires the specification of a sufficiently large constant M, which can typically be determined by an investigation of the structure of problem (9).

Alternatively, many commercial solver packages allow to directly specify the following reformulation of problem (9) via the use of piecewise linear constraints:

$$\begin{array}{rll} Z^\star_{\mathrm{ICC}} = \min\limits_{q,s,t,x} & \boldsymbol{c}^\top \boldsymbol{x} & \\ \text{s.t.} & \varepsilon N t - \boldsymbol{e}^\top \boldsymbol{s} \geq \theta N \|\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x}\| & \\ & ((\boldsymbol{b} - \boldsymbol{A}^\top \boldsymbol{x})^\top \hat{\ }_i + b - \boldsymbol{a}^\top \boldsymbol{x})^+ \geq t - s_i & \forall i \in [N] \\ & \boldsymbol{s} \geq \boldsymbol{0},\ \boldsymbol{x} \in \mathcal{X} & \end{array}$$

This formulation has the advantage that it does not require the specification of the constant M.

2.4. Reformulation of Joint Chance Constraints with Right-Hand Side Uncertainty

Assume next that problem (2) accommodates a joint chance constraint defined through the safety set $\mathcal{S}(\boldsymbol{x}) = \{\ \in \mathbb{R}^K \mid \boldsymbol{a}_m^\top \boldsymbol{x} < \boldsymbol{b}_m^\top\ + b_m\ \forall m \in [M]\}$, in which the uncertainty affects only the right-hand sides of the safety conditions. Without loss of generality, we may assume that $\boldsymbol{b}_m \neq \boldsymbol{0}$ for all $m \in [M]$. Indeed, if $\boldsymbol{b}_m = \boldsymbol{0}$, then the m^{th} safety condition in the chance constraint becomes independent of the uncertainty and can thus be absorbed in $\mathcal{X}$. Joint chance constrained programs with right-hand side uncertainty have been proposed, among others, for problems in transportation (Luedtke et al. 2010), lot-sizing (Beraldi and Ruszczyński 2002, Küçükyavuz 2012), unit commitment (Yanagisawa and Osogami 2013) and project management (Wiesemann et al. 2012).

Observe that the complement of the safety set is now representable as $\bar{\mathcal{S}}(\boldsymbol{x}) = \bigcup_{m \in [M]} \mathcal{H}_m(\boldsymbol{x})$, where $\mathcal{H}_m(\boldsymbol{x}) = \{\ \in \mathbb{R}^K \mid \boldsymbol{a}_m^\top \boldsymbol{x} \geq \boldsymbol{b}_m^\top\ + b_m\}$ is a closed halfspace for every $m \in [M]$. By Lemma 2 in the appendix we have

$$\mathbf{dist}(\hat{\ }_i, \bar{\mathcal{S}}(\boldsymbol{x})) = \min_{m \in [M]} \left\{ \frac{(\boldsymbol{b}_m^\top \hat{\ }_i + b_m - \boldsymbol{a}_m^\top \boldsymbol{x})^+}{\|\boldsymbol{b}_m\|} \right\} = \left(\min_{m \in [M]} \left\{ \frac{\boldsymbol{b}_m^\top \hat{\ }_i + b_m - \boldsymbol{a}_m^\top \boldsymbol{x}}{\|\boldsymbol{b}_m\|} \right\} \right)^+. \tag{12}$$

With this closed-form expression for the distance to the unsafe set, we can reformulate problem (2) as a mixed integer conic program.

Proposition 2. *For the safety set $\mathcal{S}(\boldsymbol{x}) = \{\ \in \mathbb{R}^K \mid \boldsymbol{a}_m^\top \boldsymbol{x} < \boldsymbol{b}_m^\top\ + b_m\ \forall m \in [M]\}$, where $\boldsymbol{b}_m \neq \boldsymbol{0}$ for all $m \in [M]$, the chance constrained program (2) is equivalent to the mixed integer conic program*

$$\begin{array}{rll} Z^\star_{\mathrm{JCC}} = \min\limits_{q,s,t,x} & \boldsymbol{c}^\top \boldsymbol{x} & \\ \text{s.t.} & \varepsilon N t - \boldsymbol{e}^\top \boldsymbol{s} \geq \theta N & \\ & \dfrac{\boldsymbol{b}_m^\top \hat{\ }_i + b_m - \boldsymbol{a}_m^\top \boldsymbol{x}}{\|\boldsymbol{b}_m\|} + \mathrm{M} q_i \geq t - s_i & \forall m \in [M],\ i \in [N] \\ & \mathrm{M}(1 - q_i) \geq t - s_i & \forall i \in [N] \\ & \boldsymbol{q} \in \{0,1\}^N,\ \boldsymbol{s} \geq \boldsymbol{0},\ \boldsymbol{x} \in \mathcal{X}, & \end{array} \tag{13}$$

where M *is a suitably large (but finite) positive constant.*

Proof of Proposition 2. By Theorem 3, the chance constrained program (2) is equivalent to (7). Using (12), the i^{th} member of the second constraint group in (7) can be reformulated as

$$\left(\min_{m2[M]} \left\{ \frac{\boldsymbol{b}_m^\top \hat{}_i + b_m - \boldsymbol{a}_m^\top \boldsymbol{x}}{\|\boldsymbol{b}_m\|} \right\} \right)^+ \geq t - s_i.$$

To eliminate the maximum operator, we introduce a binary variable $q_i \in \{0,1\}$ to re-express the above constraint as

$$\begin{cases} \dfrac{\boldsymbol{b}_m^\top \hat{}_i + b_m - \boldsymbol{a}_m^\top \boldsymbol{x}}{\|\boldsymbol{b}_m\|} + \mathrm{M} q_i \geq t - s_i \quad \forall m \in [M] \\ \mathrm{M}(1 - q_i) \geq t - s_i \end{cases}$$

A similar argument as in the proof of Proposition 1 shows that a finite value of M is sufficient for our reformulation to be exact.

Similar to Remark 5 in the previous section, many commercial solvers allow to directly specify a reformulation of problem (13) that replaces the constant M with piecewise linear constraints.

Remark 6. The deterministic reformulation (13) has N binary variables, $L+N+1$ continuous decisions as well as $(M+1)N+1$ constraints (excluding those that describe $\mathcal{X}$). In comparison, the corresponding classical chance constrained formulation has N binary variables, L continuous decisions as well as $MN+1$ constraints. Thus, adding distributional robustness requires an additional $N+1$ continuous decisions as well as N further (linear) constraints.

3. Numerical Experiments

We compare our exact reformulation of the ambiguous chance constrained program (2) with the bicriteria approximation scheme of Xie and Ahmed (2020) on a portfolio optimization problem in Section 3.1 as well as with a classical (non-ambiguous) chance constrained formulation and a Kernel density estimator based version of the ambiguous chance constrained program over a ϕ-divergence ambiguity set on a transportation problem in Section 3.2. Our goal is to investigate the computational scalability of our reformulation as well as its out-of-sample performance in a data-driven setting. All results were produced on an Intel Xeon 2.66GHz processor with 8GB memory in single-core mode using CPLEX 12.8. Following Remark 5, we avoid the specification of the constant M in our ambiguous chance constrained program through the use of piecewise linear constraints.

3.1. Portfolio Optimization

We consider a portfolio optimization problem studied by Xie and Ahmed (2020). The problem asks for the minimum-cost portfolio investment $\boldsymbol{x}$ into K assets with random returns $\tilde{\xi}_1, \ldots, \tilde{\xi}_K$ that exceeds a pre-specified target return w with high probability $1-\varepsilon$. The problem can be cast as the following instance of the ambiguous chance constrained program (2):

$$\begin{array}{ll} \min\limits_{\boldsymbol{x}} & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{s.t.} & \mathbb{P}[\tilde{}^\top \boldsymbol{x} > w] \geq 1 - \varepsilon \quad \forall \mathbb{P} \in \mathcal{F}(\theta) \\ & \boldsymbol{x} \geq \boldsymbol{0}. \end{array} \tag{14}$$

(ε, θ)	Ratio of objective values			Ratio of runtimes		
	5%	50%	95%	5%	50%	95%
$(0.05, 0.05)$	1.6	2.4	3.2	5.2	8.3	10.8
$(0.05, 0.10)$	1.9	2.9	5.0	4.9	7.7	10.6
$(0.05, 0.20)$	2.3	2.8	3.5	3.8	4.9	7.2
$(0.10, 0.05)$	1.0	1.1	1.3	7.3	10.9	13.0
$(0.10, 0.10)$	1.5	2.3	3.1	7.1	9.7	13.3
$(0.10, 0.20)$	2.1	2.7	3.9	4.2	6.2	10.1

Table 1 Objective and runtime ratios of the bicriteria approximation scheme for different values of " and . For each parameter setting, we report the 5%, 50% and 95% quantiles over 50 randomly generated instances.

We compare our exact reformulation of problem (14) with the (σ, γ)-bicriteria approximation scheme of Xie and Ahmed (2020), which produces solutions that satisfy the ambiguous chance constraint in (14) with probability $1 - \sigma\varepsilon$, $\sigma > 1$, and whose costs are guaranteed to exceed the optimal costs in (14) by a factor of at most $\gamma = \sigma/(\sigma - 1)$. Since the bicriteria approximation scheme can readily utilize support information for the random vector $\tilde{}$, we replace the ambiguity set $\mathcal{F}(\theta)$ with $\bar{\mathcal{F}}(\theta) = \mathcal{F}(\theta) \cap \{\mathbb{P} \mid \mathbb{P}[\tilde{} \in \mathbb{R}^K_+] = 1\}$ in their approach. Contrary to the experiments conducted by Xie and Ahmed (2020), we set $\sigma = 1$. This is to the disadvantage of their approach, as it does not provide any approximation guarantees in that case, but it allows us to compare the resulting portfolios as they provide the same return guarantees. For the performance of the bicriteria approximation scheme with $\sigma > 1$, we refer to Section 6.2 of Xie and Ahmed (2020).

In our numerical experiments, we consider a similar setting as Xie and Ahmed (2020). We set $K = 50$, $w = 1$ and choose the cost coefficients $c_1, \ldots, c_{50}$ uniformly at random from $\{1, \ldots, 100\}$. Each asset return $\tilde{\xi}_i$ is governed by a uniform distribution on $[0.8, 1.5]$, and we assume that $N = 100$ training samples $\hat{}_1, \ldots, \hat{}_{100}$ are available. We use the 2-norm Wasserstein ambiguity set, which implies that our exact reformulation of problem (14) is a mixed-integer second-order cone program, and set the Wasserstein radius to $\theta \in \{0.05, 0.1, 0.2\}$. The risk threshold is set to $\varepsilon \in \{0.05, 0.1\}$.

Table 1 compares the objective values and runtimes of our exact reformulation and the bicriteria approximation scheme for various combinations of the risk threshold ε and Wasserstein radius θ. The table shows that despite incorporating additional support information, the bicriteria approximation scheme determines solutions whose costs significantly exceed those of the solutions found by our exact reformulation. Perhaps more surprisingly, the bicriteria approximation scheme is also computationally more expensive. As Figure 4 shows, however, this is an artifact of the small sample

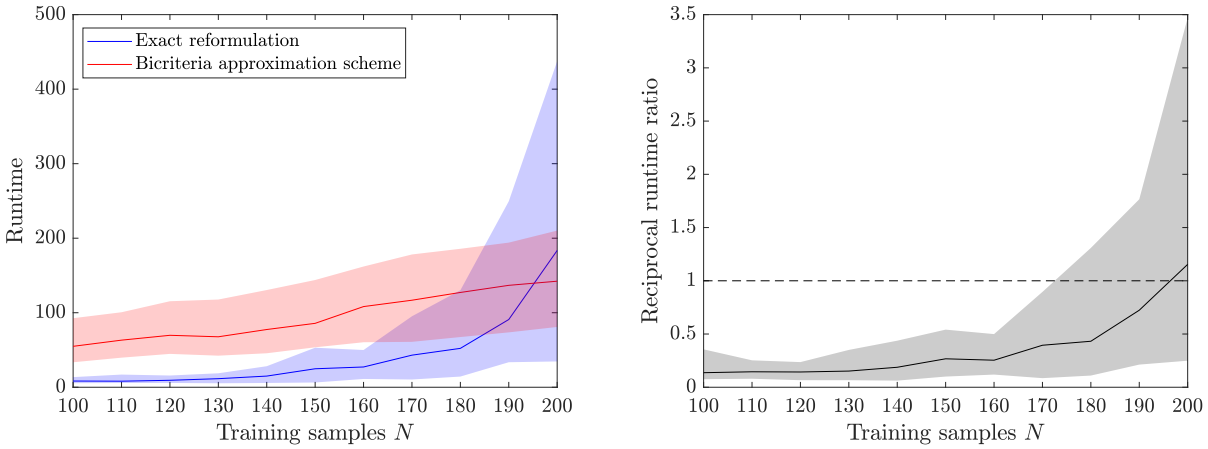

Figure 4 Runtimes (left) and reciprocal runtime ratios (right) of our exact reformulation and the bicriteria approximation scheme for $(\varepsilon; \theta) = (0.10; 0.05)$ and different sample sizes N. The shaded regions cover the 5% to 95% quantiles of 50 randomly generated instances, whereas the solid lines describe the median statistics.

size N employed in the experiments of Xie and Ahmed (2020), and the bicriteria approximation scheme is faster than our exact reformulation for larger samples sizes.

3.2. Transportation

We consider a probabilistic transportation problem studied by Luedtke et al. (2010) and Yanagisawa and Osogami (2013). The problem asks for the cost-optimal distribution of a single good from a set of factories $f \in [F]$ to a set of distribution centers $d \in [D]$. Each factory $f \in [F]$ has an individual production capacity m_f, and each distribution center $d \in [D]$ faces a random aggregate customer demand $\tilde{\xi}_d$. The cost of shipping one unit of the good from factory f to distribution center d is denoted by c_{fd}. We aim to find a transportation plan that minimizes the shipping costs, respects the production capacity of each factory and satisfies the demand at each distribution center with high probability. The problem can be cast as the following instance of problem (2):

$$\begin{array}{lll} \min\limits_{x} & \boldsymbol{c}^\top \boldsymbol{x} & \\ \text{s.t.} & \mathbb{P}\left[\sum\limits_{f \in [F]} x_{fd} \geq \tilde{\xi}_d \quad \forall d \in [D]\right] \geq 1-\varepsilon & \forall \mathbb{P} \in \mathcal{F}(\theta) \\ & \sum\limits_{d \in [D]} x_{fd} \leq m_f & \forall f \in [F] \\ & \boldsymbol{x} \geq \boldsymbol{0}. & \end{array} \tag{15}$$

Here, x_{fd} denotes the quantity shipped from factory $f \in [F]$ to distribution center $d \in [D]$. Problem (15) is an ambiguous joint chance constrained program with right-hand side uncertainty. Since

each safety condition in (15) contains a single random variable with coefficient 1 on the right-hand side, our exact reformulation reduces to the same mixed-integer linear program for any norm $\|\cdot\|$.

In our first experiment, we investigate the scalability of the exact reformulation of problem (15) that is offered by Proposition 2. To this end, we generate random test instances with 5 factories and $10, 20, \ldots, 50$ distribution centers that are located uniformly at random on the Euclidean plane $[0, 10]^2$. We identify the transportation costs c_{fd} with the Euclidean distances between the factories and distribution centers. The demand vector $\tilde{}$ is described by 50, 100 or 150 samples from a uniform distribution that is supported on [0.8 , 1.2], where the expected demand μ_d at distribution center $d \in [D]$ is picked uniformly at random from the interval $[0, 10]$. The capacity of each factory is chosen uniformly at random, and the capacities are subsequently scaled so that the factories can jointly produce up to 150% of the maximum cumulative demand. For each instance, we choose 10 ascending Wasserstein radii $\theta_1 < \ldots < \theta_{10}$ uniformly so that $\theta_1 = 0.001$ and θ_{10} is the smallest radius for which the corresponding instance of problem (15) becomes infeasible. We fix $\varepsilon = 0.1$.

Tables 2–4 and Figure 5 compare the runtimes of our ambiguous chance constrained program with those of the classical chance constrained formulation of problem (15),

$$\begin{array}{ll} \min\limits_{x,y} & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{s.t.} & \sum\limits_{f \in [F]} x_{fd} + \mathrm{M} y_i \geq \hat{\xi}_{id} \quad \forall d \in [D],\ i \in [N] \\ & \boldsymbol{e}^\top \boldsymbol{y} \leq \lfloor \varepsilon N \rfloor \\ & \sum\limits_{d \in [D]} x_{fd} \leq m_f \qquad \forall f \in [F] \\ & \boldsymbol{x} \geq \boldsymbol{0},\ \boldsymbol{y} \in \{0,1\}^N, \end{array} \tag{16}$$

where M is a sufficiently large positive constant. The results show that for the smallest Wasserstein radius $\theta_1 = 0.001$, the ambiguous chance constrained program (15) is—as expected—more difficult to solve than the corresponding classical chance constrained program (16). Interestingly, the ambiguous chance constrained program becomes considerably *easier* to solve than the classical chance constrained program for the larger Wasserstein radii $\theta_2, \ldots, \theta_{10}$. This surprising result is explained in Figure 6, which shows that the feasible region of the ambiguous chance constrained program tends to convexify as the Wasserstein radius θ increases. In fact, one can show that the set of vectors $\boldsymbol{q} \in \{0,1\}^N$ that are feasible in the deterministic reformulation of problem (15) shrinks monotonically with θ. Since it is the presence of these binary vectors that causes the non-convexity of problem (15), one can expect the problem to become better behaved as θ increases.

We next compare the out-of-sample performance of our ambiguous chance constrained program (15), where the risk threshold $\varepsilon \in \{0.1, 0.05, 0.01\}$ and the Wasserstein radius $\theta \in \{1\mathrm{E}-i : i = 2, 3, \ldots, 6\}$ are selected using a 7-fold cross-validation on the training dataset ('DRO'), with

# of distribution centers	CC	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
10	0.5	3.0	0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1
20	4.0	9.7	0.2	0.1	0.1	0.1	<0.1	<0.1	<0.1	0.1	0.1
30	7.3	13.1	0.3	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.2
40	11.2	19.3	0.4	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.3
50	15.8	166.5	0.3	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.3

Table 2 Solution times in seconds for $N = 50$ training samples. 'CC' and '$_i$' refer to problem (16) and problem (15) with different Wasserstein radii, respectively. We present median results over 100 random instances. Where the median solution time exceeds 3,600s, we report the median optimality gap in brackets.

# of distribution centers	CC	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
10	16.3	166.4	4.7	2.0	1.5	1.4	1.4	1.4	1.4	1.5	1.8
20	93.6	1910.8	8.1	2.9	2.5	2.5	2.4	2.4	2.4	2.7	2.8
30	298.3	[0.2%]	12.0	4.0	3.5	3.3	3.2	3.3	3.2	3.6	3.8
40	664.2	[0.8%]	16.0	5.1	4.7	4.5	4.5	4.5	4.4	4.8	5.1
50	1,293.2	[0.8%]	20.3	6.5	5.6	5.5	5.4	5.4	5.4	5.7	6.2

Table 3 Solution times for $N = 100$ training samples. The table has the same interpretation as Table 2.

# of distribution centers	CC	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}
10	94.6	[0.7%]	85.6	48.5	44.8	44.0	42.5	43.3	43.0	52.0	77.0
20	874.2	[1.9%]	143.9	90.5	76.3	75.6	72.8	72.5	73.2	85.7	112.4
30	[0.1%]	[3.2%]	213.8	126.4	113.0	109.5	108.9	108.8	110.3	125.4	165.1
40	[0.3%]	[3.7%]	286.8	168.2	154.2	149.1	149.3	151.7	152.1	182.8	231.5
50	[0.4%]	[3.0%]	324.6	207.0	189.3	190.9	190.0	190.4	191.8	233.0	294.4

Table 4 Solution times for $N = 150$ training samples. The table has the same interpretation as Table 2.

(i) the classical chance constrained program (16), where the risk threshold is fixed to $\varepsilon = 0.1$ ('SAA'), *(ii)* a variant of the classical chance constrained program (16), where the risk threshold $\varepsilon \in \{1\text{E}-i : i = 1, 2, \ldots, 5\} \cup \{0.05\}$ is selected using a 7-fold cross-validation on the training dataset

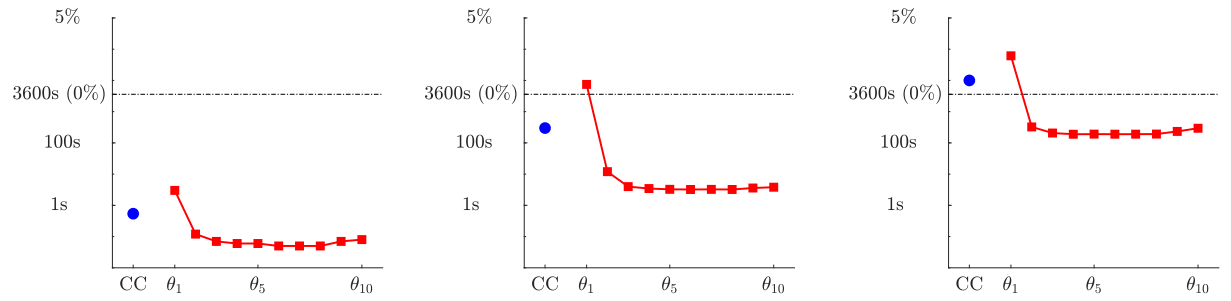

Figure 5 Median solution times (below dashed lines) and optimality gaps (above dashed lines) for $D = 10$ and $N = 50$ (left), $D = 30$ and $N = 100$ (middle) and $D = 50$ and $N = 150$ (right).

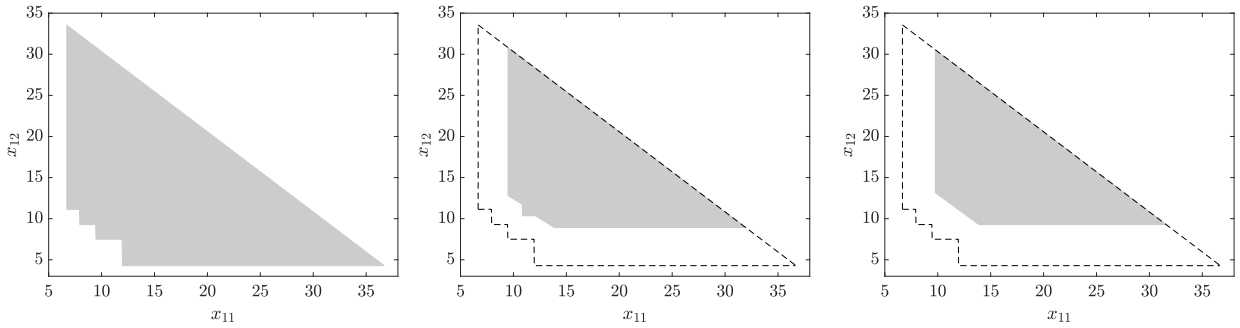

Figure 6 For a transportation problem with $F = 1$ factory, $D = 2$ distribution centers and $N = 10$ training samples, the graphs visualize the feasible regions of the classical chance constrained formulation (16) (left) and the ambiguous chance constrained problem (15) for a small (middle) and a large (right) value of .

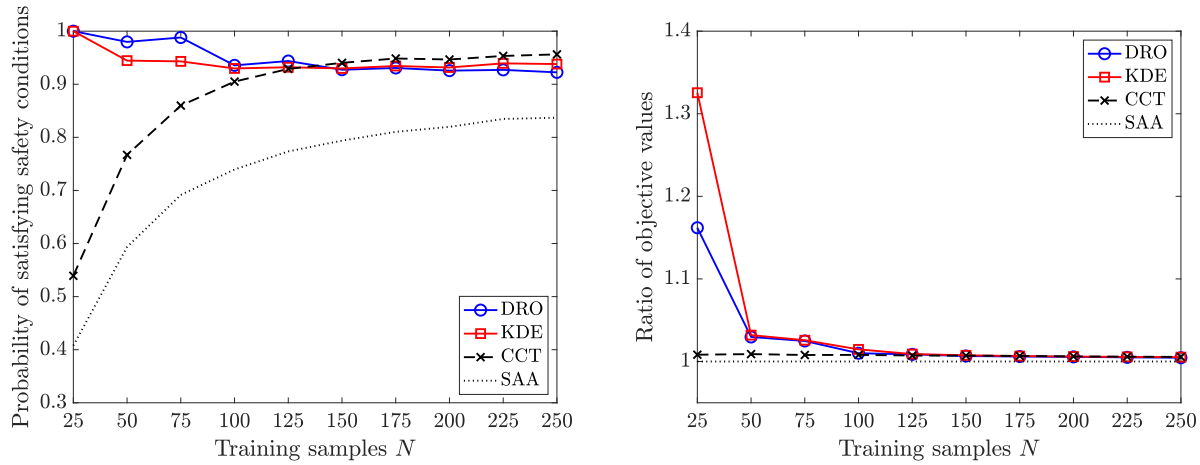

Figure 7 Probability of meeting the safety conditions (left) and transportation costs (right) for several data-driven approaches in our transportation problem with uniformly distributed demands. Both figures present median quantities over 100 random instances.

('CCT'), as well as *(iii)* a Kernel density estimator based version of the ambiguous chance constrained program over a ϕ-divergence ambiguity set, where the risk threshold $\varepsilon \in \{0.1, 0.05, 0.01\}$

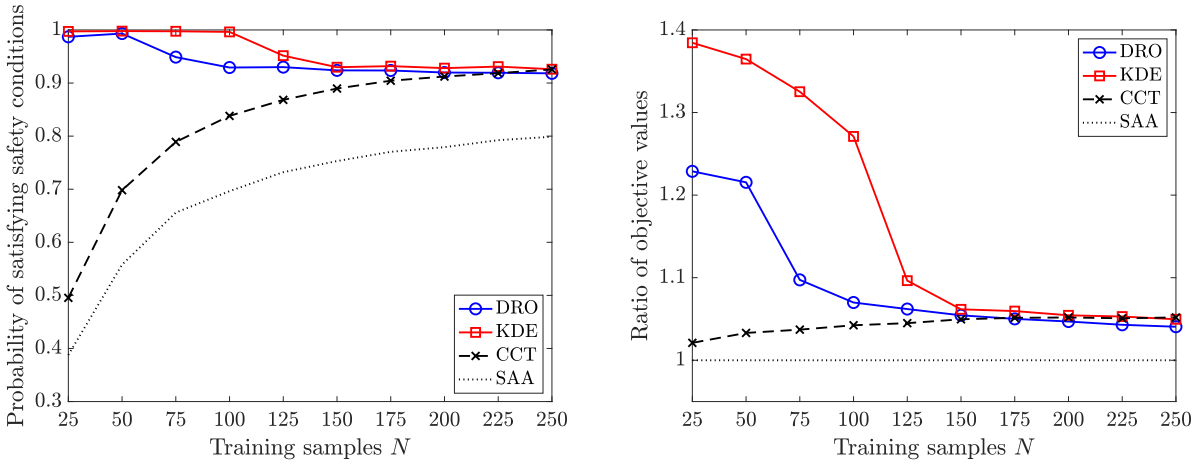

Figure 8 Probability of meeting the safety conditions (left) and transportation costs (right) for several data-driven approaches in our transportation problem with normally distributed demands. Both figures present median quantities over 100 random instances.

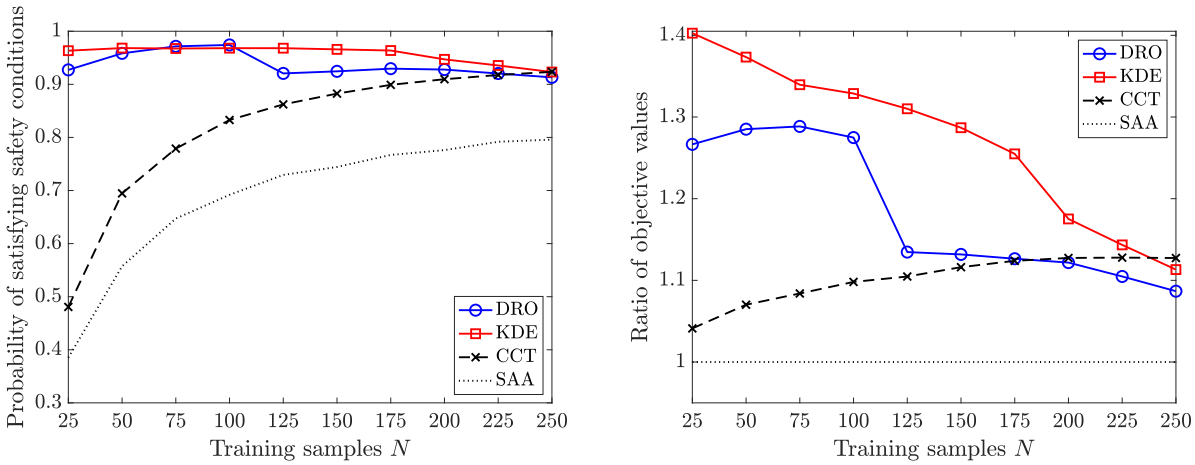

Figure 9 Probability of meeting the safety conditions (left) and transportation costs (right) for several data-driven approaches in our transportation problem with exponentially distributed demands. Both figures present median quantities over 100 random instances.

and the bandwidth $h \in \{1\mathrm{E}-i : i = -2, -1, \ldots, 3\}$ of the Gaussian kernel are selected using a 7-fold cross-validation on the training dataset ('KDE'; see Jiang and Guan 2016). We note that CCT can be regarded as a cross-validated version of the 'best data-driven reformulation' proposed by Lam (2019). We generate random problem instances with 5 factories, 20 distribution centers and 25, 30, ..., 250 training samples. In all experiments, the expected demand μ_d at distribution center $d \in [D]$ is picked uniformly at random from the interval $[0, 10]$, whereas the actual demands follow a uniform distribution that is supported on [0.8 , 1.2] (Figure 7), a normal distribution with mean and covariance matrix $0.1 \cdot \mathrm{diag}(\ )$ (Figure 8) or an exponential distribution where

each distribution center $d \in [D]$ faces a demand $(1 + 0.4 \cdot [\tilde{\zeta}_d - 0.5]) \cdot \mu_d$, where $\tilde{\zeta}_d$ follows an exponential distribution with parameter $\lambda = 2$ (Figure 9). In all cases, the demands are truncated to the non-negative real line. Our results indicate that the classical chance constrained program (16) generates solutions that significantly violate the chance constraint, even if we select the risk threshold ε out-of-sample. The two ambiguous chance constrained formulations, on the other hand, achieve the desired risk threshold, often at a modest increase in transportation costs. While our approach and the ϕ-divergence ambiguity set perform similarly, our formulation appears to result in lower transportation costs, especially when data is scarce.

Acknowledgments

The authors are grateful to the review team for constructive comments that led to substantial improvements of the paper. The authors gratefully acknowledge financial support from the ECS grant 9048191, the SNSF grant BSCGI0_157733 and the EPSRC grant EP/N020030/1.

References

Arrigo, Adriano, Christos Ordoudis, Jalal Kazempour, Zacharie De Grève, Jean-François Toubeau, François Vallée. 2022. Wasserstein distributionally robust chance-constrained optimization for energy and reserve dispatch: an exact and physically-bounded formulation. *European Journal of Operational Research* **296**(1) 304–322.

Ben-Tal, Aharon, Arkadi Nemirovski. 2001. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*. SIAM.

Beraldi, Patrizia, Andrzej Ruszczyński. 2002. A branch and bound method for stochastic integer problems under probabilistic constraints. *Optimization Methods and Software* **17**(3) 359–382.

Blanchet, Jose, Yang Kang, Karthyek Murthy. 2019. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability* **56**(3) 830–857.

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

Carlsson, John Gunnar, Mehdi Behroozi, Kresimir Mihic. 2018. Wasserstein distance and the distributionally robust TSP. *Operations Research* **66**(6) 1603–1624.

Chen, Zhi, Weijun Xie. 2021. Sharing the value-at-risk under distributional ambiguity. *Mathematical Finance* **31**(1) 531–559.

Delage, Erick, Yinyu Ye. 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research* **58**(3) 595–612.

Dert, Cees, Bart Oldenkamp. 2000. Optimal guaranteed return portfolios and the casino effect. *Operations Research* **48**(5) 768–775.

Gao, Rui, Xi Chen, Anton J Kleywegt. 2017. Distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*.

Gao, Rui, Anton J Kleywegt. 2016. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*.

Ghosal, Shubhechyya, Wolfram Wiesemann. 2020. The distributionally robust chance-constrained vehicle routing problem. *Operations Research* **68**(3) 716–732.

Goh, Joel, Melvyn Sim. 2010. Distributionally robust optimization and its tractable approximations. *Operations Research* **58**(4) 902–917.

Gounaris, Chrysanthos E, Wolfram Wiesemann, Christodoulos A Floudas. 2013. The robust capacitated vehicle routing problem under demand uncertainty. *Operations Research* **61**(3) 677–693.

Ho-Nguyen, Nam, Fatma Kılınç-Karzan, Simge Küçükyavuz, Dabeen Lee. 2020. Strong formulations for distributionally robust chance-constrained programs with left-hand side uncertainty under Wasserstein ambiguity. *arXiv preprint arXiv:2007.06750*.

Ho-Nguyen, Nam, Fatma Kılınç-Karzan, Simge Küçükyavuz, Dabeen Lee. 2021. Distributionally robust chance-constrained programs with right-hand side uncertainty under Wasserstein ambiguity. *Mathematical Programming* 1–32.

Hu, Zhaolin, Jeff Hong. 2013. Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*.

Jiang, Ruiwei, Yongpei Guan. 2016. Data-driven chance constrained stochastic program. *Mathematical Programming* **158**(1-2) 291–327.

Jiang, Ruiwei, Yongpei Guan. 2018. Risk-averse two-stage stochastic program with distributional ambiguity. *Operations Research* **66**(5) 1390–1405.

Jiang, Zhenlong, Ran Ji, Sasha Dong. 2021. A distributionally robust chance-constrained model for humanitarian relief network design. *Available at SSRN 3929286*.

Küçükyavuz, Simge. 2012. On mixing sets arising in chance-constrained programming. *Mathematical programming* **132**(1-2) 31–56.

Lam, Henry. 2019. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research* **67** 1090–1105.

Luedtke, James, Shabbir Ahmed, George L Nemhauser. 2010. An integer programming approach for linear programs with probabilistic constraints. *Mathematical Programming* **122**(2) 247–272.

Mohajerin Esfahani, Peyman, Daniel Kuhn. 2018. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming* **171**(1-2) 1–52.

Pflug, Georg, David Wozabal. 2007. Ambiguity in portfolio selection. *Quantitative Finance* **7**(4) 435–442.

Rujeerapaiboon, Napat, Daniel Kuhn, Wolfram Wiesemann. 2016. Robust growth-optimal portfolios. *Management Science* **62**(7) 2090–2109.

Serfling, Robert J. 2009. *Approximation theorems of mathematical statistics*, vol. 162. John Wiley & Sons.

Shafieezadeh-Abadeh, Soroosh, Daniel Kuhn, Peyman Mohajerin Esfahani. 2019. Regularization via mass transportation. *Journal of Machine Learning Research* **20**(103) 1–68.

Shen, Haoming, Ruiwei Jiang. 2021. Convex chance-constrained programs with Wasserstein ambiguity. *arXiv preprint arXiv:2111.02486*.

Sinha, Aman, Hongseok Namkoong, John Duchi. 2017. Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*.

Wang, Jiamin. 2007. The β-reliable median on a network with discrete probabilistic demand weights. *Operations Research* **55**(5) 966–975.

Wiesemann, Wolfram, Daniel Kuhn, Berç Rustem. 2012. Multi-resource allocation in stochastic project scheduling. *Annals of Operations Research* **193**(1) 193–220.

Wiesemann, Wolfram, Daniel Kuhn, Melvyn Sim. 2014. Distributionally robust convex optimization. *Operations Research* **62**(6) 1358–1376.

Xie, Weijun. 2019. On distributionally robust chance constrained programs with Wasserstein distance. *Mathematical Programming* 1–41.

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

Yanagisawa, H., T. Osogami. 2013. Improved integer programming approaches for chance-constrained stochastic programming. *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*. 2938–2944.

Zhang, Yiling, Jin Dong. 2021. Building load control using distributionally robust chance-constrained programs with right-hand side uncertainty and the risk-adjustable variants. *arXiv preprint arXiv:2104.11312*.

Zhao, Chaoyue, Yongpei Guan. 2018. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters* **46**(2) 262–267.

A. Distance to a Union of Halfspaces

The distance of a point $\hat{\ } \in \mathbb{R}^K$ to a closed set $\mathcal{C} \subseteq \mathbb{R}^K$ with respect to a norm $\|\cdot\|$ is defined as

$$\mathbf{dist}(\hat{\ }, \mathcal{C}) = \min\{\| \ - \ \hat{\ }\| \mid \ \in \mathcal{C}\}.$$

Note that the minimum is always attained. In the following, we derive a closed-form expression for the distance of a point to the union of finitely many closed halfspaces.

LEMMA 2. *Let $\mathcal{H}_m = \{ \ \in \mathbb{R}^K \mid a_m \geq \boldsymbol{b}_m^\top \ \}$ be a closed halfspace for each $m \in [M]$. If $\mathcal{C} = \bigcup_{m \in [M]} \mathcal{H}_m$ denotes the union of all halfspaces, then the distance of a point $\hat{\ }$ to $\mathcal{C}$ is given by*

$$\mathbf{dist}(\hat{\ }, \mathcal{C}) = \min_{m \in [M]} \left\{ \frac{(\boldsymbol{b}_m^\top \hat{\ } - a_m)^+}{\|\boldsymbol{b}_m\|} \right\} = \left(\min_{m \in [M]} \left\{ \frac{\boldsymbol{b}_m^\top \hat{\ } - a_m}{\|\boldsymbol{b}_m\|} \right\} \right)^+.$$

Proof of Lemma 2. We first prove the assertion for $M = 1$, in which case $\mathcal{C} = \mathcal{H}_1$. We thus have

$$\begin{aligned}
\mathbf{dist}(\hat{\ }, \mathcal{C}) = \ & \min_{\zeta,} \left\{ \zeta \mid \zeta \geq \| \ - \ \hat{\ }\|, \ a_1 \geq \boldsymbol{b}_1^\top \ \right\} \\
= \ & \max_{u, \boldsymbol{v}, w} \left\{ \boldsymbol{v}^\top \hat{\ } - w a_1 \ \middle| \ u = 1, \ \boldsymbol{v} = \boldsymbol{b}_1 w, \ u \geq \|\boldsymbol{v}\| \ , \ w \geq 0 \right\} \\
= \ & \max_{w} \left\{ (\boldsymbol{b}_1^\top \hat{\ } - a_1) w \ \middle| \ w \leq 1/\|\boldsymbol{b}_1\| \ , \ w \geq 0 \right\} \\
= \ & \frac{(\boldsymbol{b}_1^\top \hat{\ } - a_1)^+}{\|\boldsymbol{b}_1\|},
\end{aligned}$$

where the second equality follows from strong conic duality, which holds because the primal minimization problem is strictly feasible. Similarly, for $M \geq 1$ we find

$$\mathbf{dist}(\hat{\ }, \mathcal{C}) = \min_{m \in [M]} \mathbf{dist}(\hat{\ }, \mathcal{H}_m) = \min_{m \in [M]} \left\{ \frac{(\boldsymbol{b}_m^\top \hat{\ } - a_m)^+}{\|\boldsymbol{b}_m\|} \right\} = \left(\min_{m \in [M]} \left\{ \frac{\boldsymbol{b}_m^\top \hat{\ } - a_m}{\|\boldsymbol{b}_m\|} \right\} \right)^+,$$

where the second equality follows from the first part of the proof.