# Data-Driven Optimization with Distributionally Robust Second-Order Stochastic Dominance Constraints

Chun Peng, Erick Delage
GERAD & Department of Decision Sciences, HEC Montréal, Montréal, Canada
{chun.peng@hec.ca, erick.delage@hec.ca}

Optimization with stochastic dominance constraints has recently received an increasing amount of attention in the quantitative risk management literature. Instead of requiring that the probabilistic description of the uncertain parameters be exactly known, this paper presents the first comprehensive study of a data-driven formulation of the distributionally robust second-order stochastic dominance constrained problem (DRSSDCP) that hinges on using a type-1 Wasserstein ambiguity set. This formulation allows us to identify solutions with finite sample guarantees and solutions that are asymptotically consistent when observations are independent and identically distributed. It is furthermore for the first time shown to be axiomatically motivated in an environment with distribution ambiguity. Leveraging recent results in the field of robust optimization, we further formulate the DRSSDCP as a multistage robust optimization problem, and further propose a tractable conservative approximation that exploits finite adaptability and a scenario-based lower bounding problem, both of which can reduce to linear programs under mild conditions. We then propose the first exact optimization algorithm for this DRSSDCP, which efficiency is confirmed by our numerical results. Finally, we illustrate how the data-driven DRSSDCP can be applied in practice on resource allocation problems with both synthetic and real data. Our empirical results show that with a proper adjustment of the size of the Wasserstein ball, DRSSDCP can reach "acceptable" out-of-sample feasibility while generating strictly better performance than what is achieved by the reference strategy.

*Key words*: Robust stochastic dominance, distributionally robust optimization, Wasserstein ambiguity set, affine decision rule, exact solution algorithm, resource allocation, out-of-sample SSD feasibility

## 1. Introduction

The fundamental concept of *stochastic dominance* (or *stochastic order*) dates back to the 1940s, where it emerged in the field of statistics and economics (see an early survey by Bawa (1982)). The literature primarily focuses on two types of stochastic dominance relations: *first-order stochastic dominance* (FSD) and *second-order stochastic dominance* (SSD). Optimization with stochastic dominance constraints (SDCs) (e.g., Dentcheva and Ruszczynski 2003, Luedtke 2008, Rudolf and Ruszczyński 2008, Homem-de Mello and Mehrotra 2009, Hu et al. 2012, Haskell et al. 2017, Noyan and Rudolf 2018) has been an attractive approach to manage risk over the past decade. The classical approach minimizes some cost function $c(\boldsymbol{x})$ subject to the constraint that a controlled random performance function $f(\boldsymbol{x},\boldsymbol{\xi})$ is preferable than a given reference random performance $Y$. Here, “preferable” means that the controlled performance stochastically dominates the reference performance.

A popular example consists of a portfolio selection problem, where one wishes to choose what proportions, denoted by $\boldsymbol{x}\in\mathbb{R}^m$, of his/her capital to invest in $m$ different assets. The returns of all assets, denoted by $\boldsymbol{\xi}$, are drawn from a known distribution $\mathbb{P}$. Without loss of generality, we assume that both $\boldsymbol{x}$ and $\boldsymbol{\xi}$ are in $\mathbb{R}^m$. Let the vector $\boldsymbol{x}_0$ denote a reference portfolio, which might be a market index or an existing portfolio. The objective is for example to maximize the investment's expected return subject to the constraint that the random return stochastically dominates $\boldsymbol{\xi}^\top\boldsymbol{x}_0$ in the second order (denoted by $\succeq_{(2)}$). In this regard, one can solve the following optimization problem,

$$\underset{\boldsymbol{x}:\mathbf{1}^\top\boldsymbol{x}=1,\boldsymbol{x}\geq 0}{\text{maximize}}\ \mathbb{E}[\boldsymbol{\xi}^\top\boldsymbol{x}],\text{ s.t. }\boldsymbol{\xi}^\top\boldsymbol{x}\succeq_{(2)}\boldsymbol{\xi}^\top\boldsymbol{x}_0.$$

Methodologically speaking, most of the existing studies address the optimization with SDCs assuming that the underlying distribution is known, which gives rise to two important practical issues. First, the resolution of problems with SDCs can constitute a real computational challenge especially when the outcome space is continuous, necessitating the use of Sample Average Approximation (SAA) schemes (Kleywegt et al. 2002, Hu et al. 2012). Second, it is usually impossible for the decision-makers to exactly know the true distribution of random variables. Instead, in practice, it is more common to only have historical observations. Fortunately, these difficulties can sometimes be alleviated in stochastic programs by using the distributionally robust optimization (DRO) paradigm (e.g., Delage and Ye 2010, Wiesemann et al. 2014, Mohajerin Esfahani and Kuhn 2018), where the distribution of uncertain parameters is modeled as belonging to ambiguity sets that accounts for available distributional information. We refer the interested readers to a survey by Rahimian and Mehrotra (2019) for more recent advances in DRO.

While the concept of distributionally robust second-order stochastic dominance (DRSSD) was first introduced by Dentcheva and Ruszczyński (2010), this paper presents the first comprehensive study of a data-driven formulation of the DRSSD constrained problem (DRSSDCP) that hinges on using the Wasserstein ambiguity set proposed in Mohajerin Esfahani and Kuhn (2018). The proposed data-driven DRSSDCP will be well adapted to situations where the decision-maker only has access to a set of historical observations. Moreover, it will be controlled by an ambiguity aversion parameter $\epsilon$ (i.e., the radius of the ambiguity set), which can be used to cover a spectrum of models going from the empirical SDC problem, when $\epsilon=0$, to a distribution-free statewise SDC problem (Müller and Stoyan 2002) when $\epsilon=\infty$, which can be interpreted as a robust optimization problem.

To summarize, our contributions can be described as follows:

- From a decision theory point of view, we establish that the distributionally robust stochastic dominance constraints is the unique extension of a stochastic dominance constraint in an environment with distribution ambiguity as long as the decision maker's preference is monotone and maximally indecisive with respect to the underlying ambiguity. This provides a strong axiomatic motivation for employing distributionally robust stochastic dominance constraints in optimization models and, more generally, for any dominance constraint that employs an incomplete preference relation that are distribution-based.

- From a methodological point of view, we are the first to apply the theory presented in Mohajerin Esfahani and Kuhn (2018) to robustify SDCs. This allows us to formulate a DRSSDCP that is flexible enough to identify solutions with finite sample guarantees and that converge to the true optimal solution when observations are independent and identically distributed. While the resulting DRSSDCP appears to be generally intractable, we show that it can be reduced to a form of multi-stage robust optimization problem. Exploiting recent results in the field of robust optimization, we further propose a tractable conservative approximation and a lower bounding problem that reduce, under mild conditions, to linear programming (LP) models. Finally, we obtain analogous results for a version of the DRSSDCP that can be used when different sources of information are used for the controlled and reference performance functions.

- From an algorithmic point of view, we propose the first exact optimization solution method for a DRSSDCP that employs a type-1 Wasserstein ambiguity set. The algorithm integrates the two approximations iteratively; each iteration identifies promising modifications to tighten the approximation models using so-called active scenarios. Our numerical results provide evidence that our solution method is practicable, i.e., it can solve most instances of a data-driven DRSSDCP of reasonable size within 2 hours time limit.

- From an empirical point of view, we illustrate how the data-driven DRSSDCP can be applied in practice on resource allocation problems. Using a synthetic data generation environment where the i.i.d. assumption is satisfied, we confirm that out-of-sample SSD feasibility is improved by carefully tuning the level of robustification. We however observe for the first time in the literature that perfect out-of-sample feasibility comes at a heavy price in terms of optimality. To correct for this effect, we instead aspire to an acceptable level of feasibility that is based on the out-of-sample performance of a hypothetical controlled variable that is independently and identically distributed to the reference one. Our second set of experiments involve a portfolio optimization problem where data comes from a real stock market. Here, our empirical results show that the data-driven DRSSDCP, after being calibrated using cross-validation, can reach an acceptable level of out-of-sample SSD feasibility while generating significantly higher expected return than what is achieved by both an SAA approach and the reference portfolio, i.e. an equally weighted reference portfolio.

The remainder of this paper is organized as follows. Section 2 presents a review of relevant literature. We recall several fundamental concepts of stochastic dominance in Section 3. We present an axiomatic motivation for distributionally robust stochastic dominance in Section 4. Section 5 presents the modeling framework for optimization with DRSSD constraints under type-1 Wasserstein metric. We propose an exact iterative partitioning solution scheme for the DRSSDCP in Section 6. Section 7 presents a version of DRSSDCP where the distribution information decomposes with respect to the controlled and reference variables. We present and discuss the results of empirical experiments involving a simple resource allocation problem, with i.i.d. observations, and a portfolio optimization problem, with real market data, in Section 8. We give some concluding remarks and thoughts on future research directions in Section 9. Finally, all proofs can be found in the Appendix together with a brief list of supplementary materials.

**Notation:** We use boldface uppercase (e.g., $\boldsymbol{X}$) and lowercase (e.g., $\boldsymbol{x}$) characters to denote matrices and vectors respectively. We use $(x)^+$ to denote the positive part of $x$, i.e., $\max(0,x)$. We denote an indicator function by $\mathbf{1}\{\cdot\}$, which returns 1 if the statement inside the brackets is true, and 0 otherwise. We let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, where $\Omega$ is an outcome space, $\mathcal{F}$ the sigma algebra, and $\mathbb{P}$ is a probability measure from $\mathcal{M}(\Omega)$, the set of all probability measures on $(\Omega,\mathcal{F})$. We let $\mathcal{M}^r(\Omega)$ be a subset of $\mathcal{M}(\Omega)$ with finite $r$-th moment for $r\in[1,\infty)$. We let $\mathbb{E}_{\mathbb{P}}[\cdot]$ denote the mathematical expectation with respect to probability measure $\mathbb{P}$. We use $[N]$ to denote the set of running index $\{1,2,\cdots,N\}$. For some $\boldsymbol{\xi}\in\mathbb{R}^m$, we denote by $\delta_{\xi}$ the Dirac distribution that concentrates a unit of mass at $\boldsymbol{\xi}$. Given any norm $\|\cdot\|$ in $\mathbb{R}^m$, the dual norm is defined as $\|\boldsymbol{x}\|_* := \sup_{\boldsymbol{y}\in\mathbb{R}^m:\|\boldsymbol{y}\|\le 1} \boldsymbol{x}^\top\boldsymbol{y}$. Given a function $f:\mathbb{R}^m\to\mathbb{R}$, the conjugate function $f^*:\mathbb{R}^m\to\mathbb{R}$ is defined as $f^*(\boldsymbol{y}) := \sup_{\boldsymbol{x}\in\mathbb{R}^m} \boldsymbol{y}^\top\boldsymbol{x} - f(\boldsymbol{x})$. The support function of a set $\Xi\subseteq\mathbb{R}^m$ is denoted by $\delta(\boldsymbol{z}\mid\Xi)$ and defined as $\delta(\boldsymbol{z}\mid\Xi) := \sup_{\boldsymbol{\xi}\in\Xi}\boldsymbol{\xi}^\top\boldsymbol{z}$.

## 2. Literature Review

Optimization with SDCs was firstly introduced in the pioneering work of Dentcheva and Ruszczynski (2003), where they derive a LP reformulation. This methodological framework has been investigated in various fields, especially in portfolio selection (e.g., Roman et al. 2013, Post and Kopa 2017, Sehgal and Mehra 2020), optimal path (e.g., Zhang and Homem-de Mello 2016), power system optimization (e.g., Carrión et al. 2009), emergency medical service (EMS) location (e.g., Noyan 2010, Peng et al. 2020), as well as homeland security resource planning (e.g., Hu et al. 2011).

The literature on optimization with SDCs primarily focuses on FSD and SSD constraints. Since FSD constraints usually define a non-convex feasible set, most of the existing studies mainly focus on SSD constraints. To model such optimization problems, people usually assume that the probability measure is discrete, or can be approximated by a discrete one as is the case when using an empirical distribution (e.g., Dentcheva and Ruszczynski 2003, Luedtke 2008, Rudolf and Ruszczyński 2008, Dentcheva and Ruszczyński 2009, Lizyayev and Ruszczyński 2012, Noyan and Rudolf 2013, Ruszczyński 2013, Dentcheva and Wolfhagen 2015, Armbruster and Luedtke 2015, Dentcheva et al. 2016, Haskell et al. 2017, Noyan and Rudolf 2018, Noyan 2018, and references therein). Moreover, two-stage problems with SDCs in the second-stage are also studied (e.g., Dentcheva and Martinez 2012, Dentcheva and Wolfhagen 2016). Similarly, they still assume that the random parameters have a discrete distribution in order to derive two-stage linear formulations. Apart from the above classical version of stochastic dominance, Müller et al. (2017) and Huang et al. (2020) also propose fractional degree stochastic dominance that is defined in terms of a set of utility functions. Unfortunately, it is extremely difficult to elicit the form that this set should take in practice.

A related line of literature focuses on risk-averse optimization under the expected utility framework when there is incomplete preference information, i.e. about the utility function. For instance, Armbruster and Delage (2015) focuses on discrete outcome spaces and addresses this issue by proposing a LP reformulation that accounts for a set of pairwise comparisons. Haskell et al. (2016) extends this work by considering ambiguity about both preferences and distribution. They obtain a LP reformulation under the assumption of a polyhedral distributional ambiguity set with a finite number of vertices. For more general ambiguity sets, they propose conservative approximations that are based on reformulation-linearization techniques and semi-definite programming. Given that SSD constraints are known to be equivalent to robust expected utility constraints where the utility function belongs to the set of monotone concave functions (see Lemma 1), our DRSSDCP can be thought of as falling in this category of models. Yet, the main distinctions with this prior work are that we consider a continuous outcome space (instead of discrete), employ a Wasserstein ambiguity set, and provide an exact solution scheme for a large range of problems.

Recently, a few studies have specifically considered distributionally robust stochastic dominance constraints. Dentcheva and Ruszczyński (2010) is the first to propose a robust version of stochastic dominance for risk-averse optimization. Their work focuses on proving optimality conditions for the general formulation. Also on a more theoretical level, Chen and Jiang (2018) derives quantitative stability results for optimization problems with $k$-th order distributionally robust SDC problems with respect to simultaneous perturbation of the controlled and reference variables. Neither of the two papers studies numerical schemes for such problems. Guo et al. (2017) appears to be the first to propose numerical schemes for DRSSDCP with a moment-based ambiguity set. Their approach relies on a dense discretization of the outcome space to approximate the DRSSDCP. Unfortunately, such an approach necessarily becomes intractable as the dimension of $\boldsymbol{\xi}$ increases. In contrast, our work focuses on the Wasserstein ambiguity set, which is known to have better asymptotic properties than moment-based sets (see Theorem 3.6 in Mohajerin Esfahani and Kuhn (2018)). Furthermore, our proposed solution scheme will rely on an adaptive discretization of a one-dimensional bounded interval irrespective of the dimension of $\boldsymbol{\xi}$.

Perhaps, the closest related studies consist in Sehgal and Mehra (2020) and Kozmík (2019), although our work can be considered as independently performed as it was mostly accomplished during an overlapping period. Sehgal and Mehra (2020) studies a portfolio optimization problem with SSD constraints where the discrete return scenarios are allowed to be jointly perturbed inside a budgeted uncertainty set. The authors reformulate the problem as a LP by exploiting well-known results from Dentcheva and Ruszczynski (2003) and Bertsimas and Sim (2004). Kozmík (2019) rather studies different variations of portfolio optimization problem with DRSSD constraints where the distribution belongs to a type-1 Wasserstein ball that is intersected with different subsets of the space of $M$-points distributions. In the most general setting, they reduce the feasibility problem of DRSSDCP, i.e. identifying a worst-case $M$-point distribution for the SSD constraint, to a large non-convex polynomial optimization problem. Given the difficulty of resolution of such a model, they further restrict the ambiguity set to include information about either return scenarios or probabilities. In both cases, they identify the LP based conservative approximations.

The main distinction with these two works consists in the fact that our work is the only one that studies an untempered version of the Wasserstein ambiguity set proposed in Mohajerin Esfahani and Kuhn (2018). This is how our solutions inherit strong statistical guarantees that were derived for this type of ambiguity set and which are not known to hold for the solutions produced using the two other works. Another important distinction with Sehgal and Mehra (2020) resides in the fact that we allow the reference performance to be defined in terms of a continuous random variable with inexactly known distribution instead of an exactly known discrete distribution function. We will actually briefly discuss the case of a known discrete distribution function in Section 7, where an equivalent LP representation will also be derived for our DRSSDCP. Conversely, unlike the work of Kozmík (2019) who only identifies a tractable conservative approximation, we additionally propose an efficient exact algorithm for solving a DRSSDCP which a priori should constitute an even harder challenge given that the worst-case analysis considers a larger and less structured space of probability measures.

## 3. Preliminaries

Section 3.1 firstly recalls the basic characterizations of stochastic dominance, commonly known as the FSD, SSD, and distributionally robust stochastic dominance (DRSD). Then we present the optimization problems with DRSD in Section 3.2. We finally provide examples of how one might choose the reference performance in practice in Section 3.3. Throughout this section, we assume that $X$ captures a random revenue that needs to be maximized.

### 3.1. Stochastic Dominance

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, for any random variable $X : \Omega \to \mathbb{R}$ with distribution function $F_X^{(1)}(\eta) = \mathbb{P}(X \leq \eta)$, one can define $F_X^{(k)}(\eta) = \int_{-\infty}^{\eta} F_X^{(k-1)}(t)\,dt$, for all $\eta \in \mathbb{R}$ and $k = 2, 3, \ldots$ Hadar and Russell (1969) initially proposes the definition of stochastic dominance in the first and second order from the distribution function viewpoint.

Definition 1 (stochastic dominance). Given any two random variables $X$ and $Y$ capturing some earnings, we consider that $X$ stochastically dominates $Y$ to the $k$-th order, denoted by $X \succeq_{(k)} Y$, if and only if

$$F_X^{(k)}(\eta) \leq F_Y^{(k)}(\eta), \forall \eta \in \mathbb{R}.$$

Furthermore, the dominance is known as first-order stochastic dominance when $k = 1$ and second-order stochastic dominance when $k = 2$.

Note that SSD has been extensively studied in the literature (e.g., Dentcheva and Ruszczynski 2003, Luedtke 2008, Rudolf and Ruszczyński 2008, Homem-de Mello and Mehrotra 2009), in which a number of equivalent representations are known.

Lemma 1. *The property $X \succeq_{(2)} Y$ is equivalent to:*

1. $\mathbb{E}[(\eta - X)^+] \leq \mathbb{E}[(\eta - Y)^+], \forall \eta \in \mathbb{R}$
2. $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$ *for all non-decreasing concave utility functions $u : \mathbb{R} \to \mathbb{R}$.*

The second equivalence denoted in the lemma points to a valuable interpretation of SSD which states that $X$ is preferred to $Y$ by all risk averse expected utility maximizer's.

Note that the case where $k = 0$ also exists and is known as zero-th order stochastic dominance (or almost sure dominance), and refers to the property that:

$$\mathbb{P}(X \geq Y) = 1\,.$$

The notion of stochastic dominance needs to be modified in situations where the probability measure is undefined. In particular, one might instead need to consider a measurable space $(\Omega, \mathcal{F})$ with an ambiguous set of probability measures $\mathcal{P} \subseteq \mathcal{M}(\Omega)$, where $\mathcal{M}(\Omega)$ the set of all probability measures on $(\Omega, \mathcal{F})$. In this context, Dentcheva and Ruszczyński (2010) proposes a robust stochastic dominance criterion.

Definition 2 (Distributionally robust stochastic dominance, DRSD). Given two random variables $X$ and $Y$ and an ambiguity set $\mathcal{P}$, we say that $X$ robustly stochastically dominates $Y$ to the $k$-th order if and only if:

$$X \succeq_{(k)}^{\mathbb{P}} Y \quad \forall \mathbb{P} \in \mathcal{P},$$

where $X \succeq_{(k)}^{\mathbb{P}} Y$ refers to the fact that $X$ stochastically dominates $Y$ to the $k$-th order when the probability measure attached to $(\Omega, \mathcal{F})$ is $\mathbb{P}$. Again, we refer to this relation as *distributionally robust FSD* and *distributionally robust SSD* when $k = 1$ and $k = 2$ respectively.

### 3.2. Optimization with Stochastic Dominance Constraints

In the context of decision making and optimization, stochastic dominance can be used to ensure that the random controlled performance of one's action dominates a reference uncertain performance. Specifically, to simplify the notation, we consider the measurable space $(\Xi, \mathcal{B}(\Xi))$, with $\Xi \subseteq \mathbb{R}^m$ and $\mathcal{B}(\Xi)$ the Borel sigma algebra on $\Xi$, which assumes that all the uncertainty is captured in a vector of uncertain parameters $\boldsymbol{\xi} \in \Xi$.

Given that the probability measure on $(\Xi, \mathcal{B}(\Xi))$ is known to be $\mathbb{P}$, the $k$-th order stochastic dominance constrained optimization problem takes the form:

$$\begin{aligned} [\text{SDCP}k] \quad & \underset{\boldsymbol{x} \in \mathcal{X}}{\text{minimize}} \; h(\boldsymbol{x}) && \text{(1a)} \\ & \text{subject to } f(\boldsymbol{x}, \boldsymbol{\xi}) \succeq_{(k)}^{\mathbb{P}} f_0(\boldsymbol{\xi}) && \text{(1b)} \end{aligned}$$

where $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^m$ is the vector of decision variable in its feasible set, $h(\boldsymbol{x})$ is a cost function that needs to be minimized, $f(\boldsymbol{x}, \boldsymbol{\xi})$ is the random controlled performance function (that we wish to maximize) of $\boldsymbol{x}$ which needs to dominate the random reference performance captured by $f_0(\boldsymbol{\xi})$. Note that in practice, it can be useful to model $f_0(\boldsymbol{\xi}) := f(\boldsymbol{x}_0, \boldsymbol{\xi})$ for some reference action $\boldsymbol{x}_0 \in \mathcal{X}$.

In the case that $\mathbb{P}$ is only known to be part of $\mathcal{P}$, a distributionally robust version of problem (1) can be used.

$$\begin{aligned} [\text{DRSDCP}k] \quad & \underset{\boldsymbol{x} \in \mathcal{X}}{\text{minimize}} \; h(\boldsymbol{x}) && \text{(2a)} \\ & \text{subject to } f(\boldsymbol{x}, \boldsymbol{\xi}) \succeq_{(k)}^{\mathbb{P}} f_0(\boldsymbol{\xi}) \qquad \forall \mathbb{P} \in \mathcal{P}. && \text{(2b)} \end{aligned}$$

This paper mainly focuses on the version of problem (2) where $k = 2$ and $\mathcal{P}$ takes the form of a type-1 Wasserstein ambiguity set centered at some empirical distribution $\hat{\mathbb{P}}$.

### 3.3. Examples of Schemes for Choosing the Reference Performance

Stochastic dominance based-approach requires the decision maker to select a random reference performance that directly defines the feasible region. This reference performance therefore plays an important role on the optimality of decisions. In this section, we briefly present examples of how reference performance have been selected in different applications found in the literature. The hope is that this section can serve as inspiration for practitioners.

One common way of selecting a reference performance is to base it on the performance of a reference action, e.g., $Y := f(\boldsymbol{x}_0, \boldsymbol{\xi})$ where $\boldsymbol{x}_0$ is this reference action. For instances, in portfolio optimization, the models usually employ a reference portfolio (e.g., uniform portfolio $\boldsymbol{x}_0 = [1/n, \cdots, 1/n]$ with $n$ available assets) (e.g., Dentcheva and Ruszczynski 2003, Luedtke 2008, Guo et al. 2017, Kozmík 2019). Another example is in Peng et al. (2020), which considers a EMS location problem and where a reference coverage profile is constructed based on the solution of the deterministic formulation with nominal parameter settings. Zhang and Homem-de Mello (2017) consider minimizing the expected cost of a path while imposing a stochastic dominance constraint to control the risk of long delays. They suggest using the route obtained from a typical route-finding software as reference.

An alternative way to choose the reference performance consists in identifying a reference distribution. For instance, in portfolio optimization problems, many works have used the distribution of returns of a financial market index, which might be summarized using an empirical distribution (e.g., Roman et al. 2013, Sehgal and Mehra 2020, Liesiö et al. 2020). Noyan (2010) define a reference coverage level for EMS service based on the distribution of recent historical emergency demand. Moreover, Carrión et al. (2009) study a risk-averse electricity contract design problem with SSD constraints and similarly consider a reference cost profile defined based on a finite number of observations. A predefined distribution is further used as reference in AlAshery et al. (2019), which study a SSD constrained wind power producer's bidding problem. Finally, Zarif et al. (2012) study a mid-term scheduling problem for large industrial consumers with SSD constraints and constructs the reference performance using the distribution of the performance of optimal solutions in nominal models identified using clustering techniques.

It is worth noting that many of the prior studies consider a distribution-based reference performance rather than a reference random variable (or reference action). In a known probability setting, this choice might often be for simplicity of presentation given that the two approaches are equivalent. This is no longer the case when considering an ambiguous probability setting. While most of this paper will consider the more general case of employing a random variable as reference, Section 7 will also propose some reformulations that are specialized for distribution-based reference performances (see for instance Proposition 9).

We finally remark that depending on the choice of a reference performance, the DRSDCP$k$ might become infeasible. This can be verified by solving a phase-1 DRSDCP$k$ of the form:

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathcal{X},s}{\text{minimize}}\ s \\ &\text{subject to}\ f(\boldsymbol{x},\boldsymbol{\xi})+s\succeq_{(k)}^{\mathbb{P}} f_0(\boldsymbol{\xi}) \qquad\qquad \forall\mathbb{P}\in\mathcal{P}, \end{aligned}$$

for which a strictly positive $s^*$ indicates that the the feasible space in DRSDCP$k$ is infeasible. Moreover, the magnitude of a negative $s^*$ can provide information about the "size" of the feasible set of DRSDCP$k$. In what follows, we will in time consider that $f_0(\boldsymbol{\xi}):=f(\boldsymbol{x}_0,\boldsymbol{\xi})$, for some $\boldsymbol{x}_0\in\mathcal{X}$, which by construction implies that $\boldsymbol{x}_0$ is a feasible solution to the problem.

## 4. An Axiomatic Motivation for DRSD Constraints

Although there are a number of articles that study the robust stochastic dominance and its application in the literature (e.g., Dentcheva and Ruszczyński 2010, Chen and Jiang 2018, Sehgal and Mehra 2020), to the best of our knowledge, it is still unclear whether such type of constraints are well motivated from an axiomatic perspective. In this section, we propose a motivation for the DRSD constraint that will identify two axioms, namely "ambiguity monotonicity" and "maximally ambiguity indecisiveness", as needed to make the DRSD constraint the only possible extension of a stochastic dominance constraint in an ambiguous probability space (as defined in Delage et al. (2019)).

Formally speaking, let us consider a non-atomic ambiguous probability space $(\Omega,\mathcal{F},\mathcal{P})$ (see Definition 13 in Delage et al. (2019)). Let $\mathcal{L}_\infty(\Omega,\mathcal{F},\mathcal{P})=\cap_{\mathbb{P}\in\mathcal{P}_0}\mathcal{L}_\infty(\Omega,\mathcal{F},\mathcal{P})$ be the space of all random variables that are essentially bounded with respect to every probability measure in the ambiguity set $\mathcal{P}$. We also define the space of unambiguous random variables as

$$\mathcal{U}:=\{X\in\mathcal{L}_\infty(\Omega,\mathcal{F},\mathcal{P})\mid\exists F_X, F_X=F_X^{\mathbb{P}},\ \forall\mathbb{P}\in\mathcal{P}\}.$$

Take any preference relation $\succeq$ on random variables in $\mathcal{L}_\infty(\Omega,\mathcal{F},\mathcal{P})$. Assume that for all unambiguous random variable, the preference relation is law-invariant, i.e.

*(Law-invariance on $\mathcal{U}$) if $\{X,Y\}\subset\mathcal{U}$ and $F_X=F_Y$, then $X\sim Y$.*

Given that the ambiguous probability space is non-atomic, this means by definition that there exists a random variable $U_0\in\mathcal{U}$ that follows the standard uniform distribution under all $\mathbb{P}\in\mathcal{P}$. We can therefore overload the notation for this preference relation to apply it on distribution functions:

$$F_1\succeq F_2\Leftrightarrow F_1^{-1}(U_0)\succeq F_2^{-1}(U_0)\,,$$

where $F^{-1}(y) := \inf\{x : F(x) \geq y\}$ so that $X' := F_1^{-1}(U_0)$ is a random variable in $\mathcal{U}$ that satisfies $F_{X'}^{\mathbb{P}} = F_1$ for all $\mathbb{P} \in \mathcal{P}$.

We can now present our main representation results for a general law-invariant preference relation as follows.

Theorem 1. *If the preference relation $\succeq$ is law-invariant on $\mathcal{U}$ and satisfies:*

- *(Ambiguity Monotonicity) If $F_X^{\mathbb{P}} \succeq F_Y^{\mathbb{P}}$ for all $\mathbb{P} \in \mathcal{P}$, then $X \succeq Y$*
- *(Maximal Ambiguity Indecisiveness) If $\exists \mathbb{P} \in \mathcal{P}$ such that $F_X^{\mathbb{P}} \not\succeq F_Y^{\mathbb{P}}$, then $X \not\succeq Y$.*

*Then, for any random variables $X, Y \in \mathcal{L}_\infty(\Omega, \mathcal{F}, \mathcal{P})$, we have that $X \succeq Y$ if and only if $F_X^{\mathbb{P}} \succeq F_Y^{\mathbb{P}}$ for all $\mathbb{P} \in \mathcal{P}$. Moreover, if $\succeq$ is transitive and reflexive on $\mathcal{U}$, then it also satisfies these properties on $\mathcal{L}_\infty(\Omega, \mathcal{F}, \mathcal{P})$.*

Note that ambiguity monotonicity was already introduced in Delage et al. (2019). The maximal ambiguity indecisiveness property appears to be new and captures the fact that the decision maker becomes indecisive about the dominance between $X$ and $Y$ the moment that none of the weak dominance of $X$ on $Y$ and of $Y$ on $X$ make a unanimous consensus among the probability measures in $\mathcal{P}$. It is also interesting to observe that $\succeq$ might not be a complete preference relation in $\mathcal{L}_\infty(\Omega, \mathcal{F}, \mathcal{P})$ even when it is complete on the set $\mathcal{U}$.

The following corollary further presents our representation result to a general $k$-th order stochastic dominance relation.

Corollary 1. *Given a preference relation $\succeq$ that reduces to $\succeq_{(k)}$ when applied on the space of unambiguous random variables and is both ambiguity monotone and maximally ambiguity indecisive, then it necessarily satisfies:*

$$X \succeq Y \Leftrightarrow F_X^{\mathbb{P}} \succeq_{(k)} F_Y^{\mathbb{P}}, \forall \mathbb{P} \in \mathcal{P}. \tag{3}$$

*Furthermore, the condition that the controlled performance $f(\boldsymbol{x}, \boldsymbol{\xi})$ be preferred to a reference performance $f_0(\boldsymbol{\xi})$ according to $\succeq$ reduces to*

$$f(\boldsymbol{x}, \boldsymbol{\xi}) \succeq_{(k)}^{\mathbb{P}} f_0(\boldsymbol{\xi}), \forall \mathbb{P} \in \mathcal{P}.$$

We also remark that, based on Proposition 3.3 (i) and (iii) of Chapter 3 in Quiggin (1993), and Theorem 1, we have the following corollary that provides an alternative formulation of our axiomatic representation for the case of DRSSD.

Corollary 2. *Given preference relation $\succeq$ that satisfies the following condition:*

$$\forall X, Y \in \mathcal{U}, X \succeq Y \Leftrightarrow \exists Z \in \mathcal{U},\ \{\mathbb{E}_{\mathbb{P}}[Z|Y]\}_{\mathbb{P} \in \mathcal{P}} = \{0\}\ \&\ F_X = F_{Y+Z}$$

*and is both ambiguity monotone and maximally ambiguity indecisive, then it necessarily satisfies:*

$$X \succeq Y \Leftrightarrow F_X^{\mathbb{P}} \succeq_{(2)} F_Y^{\mathbb{P}},\, \forall \mathbb{P} \in \mathcal{P}\,.$$

*Furthermore, the condition that the controlled performance $f(x,\xi)$ be preferred to a reference performance $f_0(\xi)$ according to $\succeq$ reduces to*

$$f(x,\xi) \succeq_{(2)}^{\mathbb{P}} f_0(\xi),\, \forall \mathbb{P} \in \mathcal{P}\,.$$

We finally note that some alternative representations have been proposed in Montes et al. (2014) for extending the notion of stochastic dominance to an ambiguous probability space. Yet, all the proposed extensions either don't satisfy basic properties one would expect from preference relations in this space, namely reflexivity and transitivity, or make additional unexplained assumptions about how the ambiguity about indecisiveness should be resolved. We refer the interested readers to Appendix B.1 for a more detailed comparison.

## 5. Data-Driven DRSSDCP

In this section, we consider the distributionally robust second-order SDCP model of the form

$$\begin{aligned} \text{[DRSSDCP]}\ \ & \underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}}\ \boldsymbol{c}^\top \boldsymbol{x} & & \text{(4a)} \\ & \text{subject to}\ f(\boldsymbol{x},\boldsymbol{\xi}) \succeq_{(2)}^{\mathbb{P}} f_0(\boldsymbol{\xi}) \qquad \forall \mathbb{P}\in\mathcal{P}, & & \text{(4b)} \end{aligned}$$

where for simplicity of exposition, we focus on minimizing a simple linear cost function $\boldsymbol{c}^\top\boldsymbol{x}$. We now present the assumptions that we make on the outcome set $\Xi$ and the form of controlled performance function $f(\boldsymbol{x},\boldsymbol{\xi})$ and reference performance function $f_0(\boldsymbol{\xi})$.

Assumption 1. *The feasible set $\mathcal{X}$ is a non-empty convex set and the outcome space $\Xi$ is a non-empty compact convex set.*

Assumption 2. *The performance functions $f(\boldsymbol{x},\boldsymbol{\xi})$ and $f_0(\boldsymbol{\xi})$ are piecewise linear concave functions in both $\boldsymbol{x}$ and $\boldsymbol{\xi}$, namely, $f(\boldsymbol{x},\boldsymbol{\xi}) := \min_{n\in[N]} \boldsymbol{a}_n(\boldsymbol{x})^\top\boldsymbol{\xi} + b_n(\boldsymbol{x})$ and $f_0(\boldsymbol{\xi}) := \min_{n\in[N]} {\boldsymbol{a}_n^0}^\top\boldsymbol{\xi} + b_n^0$ with $\boldsymbol{a}_n(\boldsymbol{x})$ and $b_n(\boldsymbol{x})$ affine in $\boldsymbol{x}$ for all $n\in[N]$.*

First note that Assumption 1 is relatively weak and often satisfied. Assumption 2 is comparatively much more restrictive, yet it is satisfied by linear performance functions, or performance functions that compute the amount of shortfall for reaching a prescribed target, e.g. $f(\boldsymbol{x},\boldsymbol{\xi}) := -\max(b - \boldsymbol{a}(\boldsymbol{\xi})^T\boldsymbol{x}, 0)$, where $\boldsymbol{a}(\boldsymbol{\xi})^T\boldsymbol{x}$, with $\boldsymbol{a}(\boldsymbol{\xi})$ affine in $\boldsymbol{\xi}$, computes the performance achieved and $b$ is the target level. More generally speaking, $f(\boldsymbol{x},\boldsymbol{\xi})$ could be the optimal value of a linear recourse optimization model with right-hand-side uncertainty:

$$f(\boldsymbol{x},\boldsymbol{\xi}) := \max_{\boldsymbol{y}:A\boldsymbol{x}+B\boldsymbol{y}\le D\boldsymbol{\xi}} \boldsymbol{d}^\top\boldsymbol{y} = \min_{n\in[N]} \boldsymbol{s}_n^\top(D\boldsymbol{\xi} - A\boldsymbol{x})\,,$$

where $\{\boldsymbol{s}_n\}_{n\in[N]}$ is the set of vertices of the feasible set of the dual linear program. However, our proposed approach will for simplicity assume that $N$ is of reasonable size.

In the following, we give the definition of Wasserstein metric and Wasserstein ambiguity set.

Definition 3 (Wasserstein Metric). For any $r \geq 1$, let $\mathcal{M}^r(\Xi)$ denote the set of all probability measures $\mathbb{P}$ on $(\Xi, \mathcal{B}(\Xi))$ satisfying $\mathbb{E}_{\mathbb{P}}[d(\boldsymbol{\xi}, \boldsymbol{\xi}_{\mathbf{0}})^r] = \int_{\Xi^2} d(\boldsymbol{\xi}, \boldsymbol{\xi}_{\mathbf{0}})^r \mathbb{P}(d\boldsymbol{\xi}) < \infty$ for the same reference point $\boldsymbol{\xi}_{\mathbf{0}} \in \Xi$ and where $d(\boldsymbol{\xi}, \boldsymbol{\xi}_{\mathbf{0}})$ is a continuous reference metric on $\Xi$. The type-$r$ Wasserstein distance between distributions $\mathbb{P}_1 \in \mathcal{M}^r(\Xi)$ and $\mathbb{P}_2 \in \mathcal{M}^r(\Xi)$ is defined as

$$d_{\mathrm{W}}^r(\mathbb{P}_1, \mathbb{P}_2) = \inf_{\mathbb{Q}\in\mathcal{M}(\mathbb{P}_1,\mathbb{P}_2)} \left( \int_{\Xi^2} d(\boldsymbol{\xi}_1, \boldsymbol{\xi}_{\mathbf{2}})^r \mathbb{Q}(d\boldsymbol{\xi}_1, d\boldsymbol{\xi}_2) \right)^{\frac{1}{r}},$$

where $\mathcal{M}(\mathbb{P}_1, \mathbb{P}_2)$ is the set of joint distribution $\mathbb{Q} \in \mathcal{M}^r(\Xi \times \Xi)$ of $\boldsymbol{\xi}_1 \in \Xi$ and $\boldsymbol{\xi}_2 \in \Xi$ with the marginal distributions equal to $\mathbb{P}_1$ and $\mathbb{P}_2$, i.e., $\mathbb{Q}(\Xi' \times \Xi) = \mathbb{P}_1(\Xi')$ and $\mathbb{Q}(\Xi \times \Xi') = \mathbb{P}_2(\Xi')$ for all $\Xi' \subseteq \Xi$.

Definition 4 (Wasserstein Ambiguity Set). The Wasserstein ambiguity set of radius $\epsilon$ centered at $\bar{\mathbb{P}}$ is defined by

$$\mathcal{P}_{\mathrm{W}}^r(\bar{\mathbb{P}}, \epsilon) := \left\{ \mathbb{P} \in \mathcal{M}^r(\Xi) \,\middle|\, d_{\mathrm{W}}^r(\mathbb{P}, \bar{\mathbb{P}}) \leq \epsilon \right\},$$

where $d_{\mathrm{W}}$ is the Wasserstein metric that is given in Definition 3.

One can think of the Wasserstein radius $\epsilon$ as a budget on the transportation cost induced by rearranging the reference distribution $\bar{\mathbb{P}}$ to obtain $\mathbb{P}$. When the only information about the true $\mathbb{P}$ consists in a limited number of sampled observations $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^M$, a natural choice for $\bar{\mathbb{P}}$ consists in the empirical distribution $\hat{\mathbb{P}} := \frac{1}{M} \sum_{i\in[M]} \delta_{\hat{\xi}_i}$, which assigns equal weights to each observed realizations. A special case of Wasserstein ambiguity set is characterized by the following assumption which will help identify a linear programming reformulation of Wasserstein-based distributionally robust stochastic dominance constraints.

Assumption 3. *The Wasserstein ambiguity set $\mathcal{P}_W^1$ uses a type-1 Wasserstein distance with the $\ell_1$-norm or $\ell_\infty$-norm as the reference metric, i.e., $d(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) := \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|_p$ with $p \in \{1, \infty\}$.*

Remark 1. Although the rest of the paper mainly focuses on $\mathcal{P}_{\mathrm{W}}^1(\hat{\mathbb{P}}, \epsilon)$ under Assumption 3, we can also straightforwardly extend the proposed solution scheme to $\mathcal{P}_{\mathrm{W}}^\infty(\hat{\mathbb{P}}, \epsilon)$ with the $\ell_1$-norm or $\ell_\infty$-norm as the reference metric. We refer the interested readers to Appendix B.2 for more details.

To provide more intuition about DRSSDCP with a Wasserstein ambiguity set, in the following we start by demonstrating in Section 5.1 that DRSSDCP (4) generalizes two popular classes of stochastic dominance problems. Section 5.2 presents statistical properties of DRSSDCP (4) that emerge when the historical observations are drawn independently and identically. We then derive a multistage robust optimization reformulation for DRSSDCP under mild conditions in Section 5.3. We then conservatively approximate it as an adaptive robust linear optimization problem by using finite adaptability and affine decision rules in Section 5.4. Finally, we provide a lower bounding approximation by using finite scenarios and derive the tractable reformulation in Section 5.5.

## 5.1. Reduction to Classical Optimization Problems

In the following two propositions, we show that, under a Wasserstein ambiguity set, the distributionally robust SSD constraint can model a wide spectrum of ambiguity aversions: ranging from the classical empirical SSD constraints when $\epsilon = 0$ to a distribution-free statewise dominance constraint when $\epsilon = \infty$, which can be interpreted as a robust optimization problem.

Proposition 1 **(Reduction to SDCP2)**. *The DRSSDCP* (4) *with* $\mathcal{P}^r_W(\hat{\mathbb{P}}, 0)$ *reduces to a SDCP2 model* (1) *with* $h(\boldsymbol{x}) := \boldsymbol{c}^\top \boldsymbol{x}$ *and* $\mathbb{P} := \hat{\mathbb{P}}$*. Moreover, under assumptions 1 and 2, it can be reformulated as a linear programming problem when* $\mathcal{X}$ *is polyhedral.*

Proposition 2 **(Reduction to DFSDCP)**. *The DRSSDCP* (4) *with* $\mathcal{P}^r_W(\hat{\mathbb{P}}, \infty)$ *reduces to the following Distribution-Free Statewise Dominance Constrained Problem (DFSDCP),*

$$\text{[DFSDCP]}\ \underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}}\ \boldsymbol{c}^\top \boldsymbol{x} \tag{5a}$$

$$\text{subject to}\ f(\boldsymbol{x},\boldsymbol{\xi}) \geq f_0(\boldsymbol{\xi}) \qquad \forall \boldsymbol{\xi} \in \Xi. \tag{5b}$$

*Moreover, under assumptions 1 and 2, problem* (5) *can be reformulated as a linear programming problem when* $\Xi$ *and* $\mathcal{X}$ *are polyhedral.*

In the remaining of this paper, we focus on the case where $\epsilon \in (0, +\infty)$ and derive reformulations and propose solution schemes for the case where a type-1 Wasserstein metric is used and either $\ell_1$-norm or $\ell_\infty$-norm is used as the reference distance $d(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$.

## 5.2. Statistical Properties of DRSSDCP Solutions

We briefly summarize some valuable properties that are held by solutions of the DRSSDCP in a data-driven context. First, Theorem 3.5 in Mohajerin Esfahani and Kuhn (2018) actually established conditions under which $\mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}}, \epsilon)$ is known to contain the true distribution that generated the i.i.d. observations $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^M$ with high probability. These conditions straightforwardly imply a finite sample guarantee for DRSSDCP solutions.

Proposition 3 **(Finite sample guarantee of DRSSDCP solutions)**. *Suppose that Assumption 1 holds and that each observations in* $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^M$ *are drawn i.i.d. from some* $\bar{\mathbb{P}}$*, with* $M \geq 1$ *and*

*$m>2$. Given some $\beta\in(0,1)$, let $\hat{\boldsymbol{x}}_M$ be the optimal solution of the DRSSDCP with ambiguity set $\mathcal{P}_W^1(\hat{\mathbb{P}},\epsilon_M(\beta))$ where*

$$\epsilon_M(\beta):=\begin{cases}\left(\dfrac{log(c_1\beta^{-1})}{c_2M}\right)^{1/\max(m,2)} & \textit{if } M\geq\dfrac{log(c_1\beta^{-1})}{c_2}\\ \left(\dfrac{log(c_1\beta^{-1})}{c_2M}\right)^{1/a} & \textit{otherwise},\end{cases}$$

*and where $c_1$, $c_2$, and $a>1$ are positive constants (see Mohajerin Esfahani and Kuhn (2018) for details). One has the guarantee that, with probability larger than $1-\beta$, $\hat{\boldsymbol{x}}_M$ satisfies the SSD constraint under $\bar{\mathbb{P}}$, i.e., $f(\hat{\boldsymbol{x}}_M,\boldsymbol{\xi})\succeq_{(2)}^{\bar{\mathbb{P}}}f_0(\boldsymbol{\xi})$.*

Alternatively, Theorem 3.1 of Hu et al. (2012) establishes conditions under which the SAA of SDCP2 (i.e., $\epsilon=0$) converges to the set of the “near-optimal solutions” as the number of samples goes to infinity. The following proposition further generalizes this result for the case when $\epsilon>0$.

Proposition 4 **(Asymptotic consistency of DRSSDCP solutions)**. *Suppose that assumptions 1 and 2 hold, that $\mathcal{X}$ is bounded, and that $\beta_M\in(0,1)$ satisfies $\sum_{M=1}^{\infty}\beta_M<\infty$ and $\lim_{M\to\infty}\epsilon_M(\beta_M)=0$. Consider the following $\phi$-SDCP2 model:*

$$[\phi\text{-}SDCP2]\quad \underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}}\ \boldsymbol{c}^\top\boldsymbol{x}$$
$$\text{subject to } \mathbb{E}_{\bar{\mathbb{P}}}\left[(t-f(\boldsymbol{x},\boldsymbol{\xi}))^+\right]\leq\mathbb{E}_{\bar{\mathbb{P}}}\left[(t-f_0(\boldsymbol{\xi}))^+\right]+\phi\qquad\forall t\in\mathbb{R},$$

*with $\phi>0$, and assume that Slater's condition is satisfied. Let each observations in $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^M$ be drawn i.i.d. from some $\bar{\mathbb{P}}$, $\boldsymbol{x}_M$ be an optimal solution of the $\phi$-DRSSDCP:*

$$[\phi\text{-}DRSSDCP]\quad \underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}}\ \boldsymbol{c}^\top\boldsymbol{x}$$
$$\text{subject to } \mathbb{E}_{\mathbb{P}}\left[(t-f(\boldsymbol{x},\boldsymbol{\xi}))^+\right]\leq\mathbb{E}_{\mathbb{P}}\left[(t-f_0(\boldsymbol{\xi}))^+\right]+\phi\quad\forall t\in\mathbb{R},\forall\mathbb{P}\in\mathcal{P}_W^1(\hat{\mathbb{P}},\epsilon),$$

*with ambiguity set $\mathcal{P}_W^1(\hat{\mathbb{P}},\epsilon_M(\beta_M))$, and $\mathcal{X}^*$ be the set of optimal solutions to the $\phi$-SDCP2 under the true distribution $\bar{\mathbb{P}}$. Then one has the guarantee that $\boldsymbol{x}_M$ converges almost surely to $\mathcal{X}^*$ as $M$ goes to infinity.*

In summary, we see that for appropriately chosen values of $\epsilon$, a slightly perturbed version of DRSSDCP can identify solutions that either achieve relevant finite sample guarantees for small observation sets, or achieve near optimality for the true underlying SDCP2 when $M$ is sufficiently large. In practice, one should rely on cross-validation schemes (e.g., as is described in Section 8.2.3) to identify the size of $\epsilon$ that is most appropriate for the data set.

### 5.3. Exact Multistage Robust Optimization Reformulation

We now focus on converting the DRSSDCP to a multistage robust optimization model using the results of Wasserstein DRO from Mohajerin Esfahani and Kuhn (2018) when Assumption 3 is satisfied. In particular, based on Lemma 1, we can first rewrite the equivalent representation of model (4) in the form of

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}}\ \boldsymbol{c}^\top \boldsymbol{x} && && \text{(6a)}\\ &\text{subject to } \mathbb{E}_{\mathbb{P}}\left[(t-f(\boldsymbol{x},\boldsymbol{\xi}))^+\right] \le \mathbb{E}_{\mathbb{P}}\left[(t-f_0(\boldsymbol{\xi}))^+\right] && \forall\, t\in\mathbb{R},\ \forall\mathbb{P}\in\mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}},\epsilon). && \text{(6b)}\end{aligned}$$

Moreover, constraint (6b) can be rewritten as

$$\sup_{\mathbb{P}\in\mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}},\epsilon)} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x},\boldsymbol{\xi},t)\right]\le 0, \quad \forall t\in\mathbb{R}\,, \tag{7}$$

where $g(\boldsymbol{x},\boldsymbol{\xi},t) := (t-f(\boldsymbol{x},\boldsymbol{\xi}))^+ - (t-f_0(\boldsymbol{\xi}))^+$. The theory of Mohajerin Esfahani and Kuhn (2018) can therefore be applied to obtain the reformulation presented in the following proposition.

Proposition 5. *Under assumptions 1 and 2, and with $\epsilon\in(0,\infty)$, the DRSSDCP* (6) *coincides with the optimal value of the following multistage robust optimization problem with two stages of decisions (i.e. $\boldsymbol{x}$ followed with $\lambda$ and $\boldsymbol{q}$) and two stages of adversarial perturbations (i.e. $t$ then $\boldsymbol{\xi}$):*

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}}\ \boldsymbol{c}^\top \boldsymbol{x} && && \text{(8a)}\\ &\text{subject to } L(\boldsymbol{x},t)\le 0 && \forall t\in\bar{\mathcal{T}}, && \text{(8b)}\end{aligned}$$

*where $\bar{\mathcal{T}} := [t_{min}, t_{max}]$, with $t_{min} := \inf_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{\xi})$, $t_{max} := \sup_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{\xi})$, and where*

$$\begin{aligned} L(\boldsymbol{x},t) :=\quad & \inf_{\lambda,\boldsymbol{q}}\ \lambda\epsilon + \frac{1}{M}\sum_{i\in[M]} q_i && && \text{(9a)}\\ &\text{subject to}\ \ g_n(\boldsymbol{x},\boldsymbol{\xi},t) - \lambda\|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_i\| \le q_i && \forall\boldsymbol{\xi}\in\Xi, n\in[N], i\in[M] && \text{(9b)}\\ &\qquad\qquad\ \ \lambda\ge 0, \boldsymbol{q}\in\mathbb{R}^M, && && \text{(9c)}\end{aligned}$$

*where*

$$g_n(\boldsymbol{x},\boldsymbol{\xi},t) := \min_{n'\in[N+1]} (\boldsymbol{a}^0_{n'} - \boldsymbol{a}_n(\boldsymbol{x}))^\top \boldsymbol{\xi} + b^0_{n'} - b_n(\boldsymbol{x}) - (c^0_{n'} - c_n)t$$

*with $a_{N+1} = b_{N+1} = c_{N+1} = a^0_{N+1} = b^0_{N+1} = c_{N+1} = 0$, and $c_n = c^0_n = 1$ for all $n\in[N]$. Moreover, it can be reformulated as a multistage robust linear optimization problem when Assumption 3 is satisfied and when $\mathcal{X}$ and $\Xi$ are polyhedral.*

We wish to emphasize the fact that most of the previously proposed solution schemes for the optimization problems with SSD constraints (e.g., Dentcheva and Ruszczynski 2003) exploit the property that when the constraint is violated, it is necessarily violated for $t$ at one of the support points of $f_0(\boldsymbol{\xi})$. The fact that $\mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}},\epsilon)$ includes distributions that make $f_0(\boldsymbol{\xi})$ continuously supported prevents us from restricting $t$ to take values in a finite set. We also note that the multistage robust optimization problem (8) takes the form of $\min\limits_{\boldsymbol{x}}$-$\sup\limits_{t}$-$\min\limits_{\boldsymbol{\lambda},\boldsymbol{q}}$-$\sup\limits_{\boldsymbol{\xi}}$, which cannot be tractably reformulated using duality theory. Next, we propose two tractable approximations that can be used to bound the optimal value of problem (8). This in turn will motivate an exact iterative partitioning optimization solution scheme for the problem.

Remark 2. More recently, there are several commonly used methods for the risk-averse DRO problems with a Wasserstein ambiguity set in the literature, e.g., distributionally robust chance constraints (e.g., Chen et al. 2018, Xie 2021), distributionally robust risk measures (e.g., Guo and Xu 2019, Ji and Lejeune 2021) and distributionally robust expected utility (e.g., Gao and Kleywegt 2016, Zhao and Guan 2018, Mohajerin Esfahani and Kuhn 2018, Long et al. 2021). The resolution methods for these risk-averse Wasserstein DRO problems usually relies on the approximation schemes (e.g., CVaR and Bonferroni approximations), the Fenchel Robust Counterpart theory (Ben-Tal et al. 2015) and strong duality to derive the tractable finite-dimensional convex reformulations under mild conditions, which can usually be solved by the state-of-art solvers (e.g., CPLEX, GUROBI) for the small/medium-sized problems. While the methods that we will employ have been used in other settings, this paper is the first to propose an exact solution scheme for risk-averse optimization problems with DRSSD constraints.

## 5.4. Tractable Conservative Approximation Formulation via Finite Adaptability

In this section, we first present a conservative approximation model of multistage robust optimization problem (8) by applying finite adaptability and then derive its tractable reformulation under mild conditions, which provides an upper bound for problem (8).

Since robust multistage optimization problems are generally hard to solve, a common approach is to employ affine decision rules to obtain a conservative approximation (Ben-Tal et al. 2004). Unfortunately, in problem (8), the term $\lambda(t)\|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_i\|$ makes this approach challenging. Therefore, we hereby consider the adjustable decisions $\lambda(t)$ and $\boldsymbol{q}(t)$ to be respectively piecewise constant and piecewise linear on a partition $\mathscr{P}:=\{\mathcal{T}_k\}_{k=1}^K$ of $\bar{\mathcal{T}}$, i.e., $\lambda(t)=\sum\limits_{k\in[K]}\lambda_k\mathbf{1}\{t\in\mathcal{T}_k\}$ and $q_i(t)=\sum\limits_{k\in[K]}(\bar{q}_{ik}+q_{ik}t)\mathbf{1}\{t\in\mathcal{T}_k\}$ respectively.

In doing so, this gives rise to the following robust optimization problem,

$$\underset{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\lambda},\boldsymbol{q},\bar{\boldsymbol{q}}}{\text{minimize}}\ \boldsymbol{c}^\top\boldsymbol{x} \tag{10a}$$

$$
\begin{aligned}
\text{subject to } & \lambda_k \epsilon + \frac{1}{M} \sum_{i \in [M]} (\bar{q}_{ik} + q_{ik} t) \le 0 && \forall t \in \mathcal{T}_k, k \in [K] && \text{(10b)} \\
& g_n(\boldsymbol{x}, \boldsymbol{\xi}, t) - \lambda_k \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i\| \le \bar{q}_{ik} + q_{ik} t && \forall \boldsymbol{\xi} \in \Xi, t \in \mathcal{T}_k, k \in [K], i \in [M], n \in [N+1] && \text{(10c)} \\
& \lambda_k \ge 0, q_{ik}, \bar{q}_{ik} \in \mathbb{R} && \forall k \in [K], i \in [M], && \text{(10d)}
\end{aligned}
$$

where $g_n(\boldsymbol{x}, \boldsymbol{\xi}, t) = \max\limits_{\eta \in \Gamma_\eta(\boldsymbol{\xi}, t)} -\boldsymbol{a}_n(\boldsymbol{x})^\top \boldsymbol{\xi} - b_n(\boldsymbol{x}) + c_n t + \eta$ with $a_{N+1} = b_{N+1} = c_{N+1} = 0$, and $c_n = 1$ for all $n \in [N]$, and $\Gamma_\eta(\boldsymbol{\xi}, t) := \left\{ \eta : \eta \le {\boldsymbol{a}^0_{n'}}^\top \boldsymbol{\xi} + b^0_{n'} - t, \ \forall n' \in [N]; \ \eta \le 0 \right\}$. Note that model (10) always constructs a conservative approximation for problem (8), as the former restricts the space of decision rules to those which are piecewise linear and static over the partition. That is to say, all feasible solutions of problem (10) are necessarily feasible in problem (8) and the optimal value of problem (10) provides a upper bound on the objective value of problem (10)'s solution in problem (8). In the following theorem, we provide an exact finite-dimensional convex optimization reformulation of model (10), in which we employ the Fenchel Robust Counterpart theory in Ben-Tal et al. (2015) to derive the equivalent reformulations of robust constraints (10b) and (10c) respectively.

THEOREM 2. *Suppose that assumptions 1 and 2 hold, for the given partition $\mathscr{P} := \{\mathcal{T}_k\}_{k=1}^K$, the conservative approximation model* (10) *is equivalent to the following finite-dimensional convex optimization problem,*

$$
\begin{aligned}
\underset{\boldsymbol{x} \in \mathcal{X}, \boldsymbol{\lambda}, \boldsymbol{\rho}, \boldsymbol{q}, \bar{\boldsymbol{q}}, \boldsymbol{v}, \boldsymbol{u}, \boldsymbol{w}}{\text{minimize}} \ & \boldsymbol{c}^\top \boldsymbol{x} \\
\text{subject to } & \lambda_k \epsilon + \frac{1}{M} \sum_{i=1}^{M} (q_{ik} \bar{t}_k^+ + \bar{q}_{ik}) \le 0, && \forall k \in [K] \\
& \lambda_k \epsilon + \frac{1}{M} \sum_{i=1}^{M} (q_{ik} \bar{t}_k^- + \bar{q}_{ik}) \le 0 && \forall k \in [K] \\
& \delta(\boldsymbol{v}_{ink} \mid \Xi) + u_{ink} \bar{t}_k^- - \boldsymbol{w}_{ink}^\top \hat{\boldsymbol{\xi}}_i - \bar{q}_{ik} \\
& - b_n(\boldsymbol{x}) + \sum_{n' \in [N]} \rho_{inkn'} b^0_{n'} \le 0 && \forall i \in [M], n \in [N+1], k \in [K] \\
& \delta(\boldsymbol{v}_{ink} \mid \Xi) + u_{ink} \bar{t}_k^+ - \boldsymbol{w}_{ink}^\top \hat{\boldsymbol{\xi}}_i - \bar{q}_{ik} \\
& - b_n(\boldsymbol{x}) + \sum_{n' \in [N]} \rho_{inkn'} b^0_{n'} \le 0 && \forall i \in [M], n \in [N+1], k \in [K] \\
& \boldsymbol{w}_{ink} = \boldsymbol{v}_{ink} + \boldsymbol{a}_n(\boldsymbol{x}) - \sum_{n' \in [N]} \rho_{inkn'} \boldsymbol{a}^0_{n'} \\
& \|\boldsymbol{w}_{ink}\|_* \le \lambda_k && \forall i \in [M], n \in [N+1], k \in [K] \\
& u_{ink} + q_{ik} + \sum_{n' \in [N]} \rho_{inkn'} - c_n = 0 && \forall i \in [M], n \in [N+1], k \in [K] \\
& \sum_{n' \in [N]} \rho_{inkn'} \le 1 && \forall i \in [M], n \in [N+1], k \in [K]
\end{aligned}
$$

$$\lambda_k, \rho_{inkn'} \geq 0; \boldsymbol{w}_{ink}, \boldsymbol{v}_{ink} \in \mathbb{R}^m; u_{ink}, \bar{q}_{ik}, q_{ik} \in \mathbb{R} \quad \forall i \in [M], n' \in [N], n \in [N+1], k \in [K],$$

*where $\bar{t}_k^-$ and $\bar{t}_k^+$ are the two boundaries of each interval $\mathcal{T}_k$. Furthermore, it can be reformulated as a linear programming problem if $\mathcal{X}$ and $\Xi$ are polyhedral and Assumption 3 is satisfied.*

### 5.5. Tractable Lower Bounding Approximation via Finite Scenarios

Since the conservative approximation model (10) becomes difficult to solve when the size of the partition becomes large, it is useful to know how far the current optimal solution is from being optimal. In this regard, we now propose a tractable approximation for problem (8) that will provide a lower bound. More specifically, inspired by the scenario-based two-stage robust optimization problem in Hadjiyiannis et al. (2011), we employ a finite scenarios set $\hat{\mathcal{T}} := \{\hat{t}_1, \cdots, \hat{t}_k, \cdots, \hat{t}_K\}$ to replace $\bar{\mathcal{T}}$ in problem (8) (e.g. uniformly spread scenarios on the interval $\bar{\mathcal{T}}$). This gives rise to the following optimization problem,

$$\begin{aligned}
&\underset{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\lambda},\boldsymbol{q}}{\text{minimize}} \quad \boldsymbol{c}^\top \boldsymbol{x} && && \text{(11a)}\\
&\text{subject to } \lambda_k \epsilon + \frac{1}{M}\sum_{i\in[M]} q_{ik} \leq 0 && \forall k \in [K] && \text{(11b)}\\
&\qquad g_n(\boldsymbol{x}, \boldsymbol{\xi}, \hat{t}_k) - \lambda_k \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i\| \leq q_{ik} && \forall \boldsymbol{\xi} \in \Xi, i \in [M], n \in [N+1], k \in [K] && \text{(11c)}\\
&\qquad \lambda_k \geq 0, q_{ik} \in \mathbb{R} && \forall i \in [M], k \in [K]. && \text{(11d)}
\end{aligned}$$

In the following Proposition 6, we show that the objective value of problem (11) provides a lower bound, and can be reformulated as a linear programming problem under mild conditions. The latter follows from employing Fenchel Robust Counterpart approaches presented in Ben-Tal et al. (2015).

Proposition 6. *Given a finite scenarios set $\hat{\mathcal{T}} = \{\hat{t}_1, \cdots, \hat{t}_k, \cdots, \hat{t}_K\}$, problem* (11) *provides a lower bounding approximation for problem* (8)*. Suppose that assumptions 1 and 2 hold, problem* (11) *is equivalent to the following finite-dimensional convex optimization problem,*

$$\begin{aligned}
&\underset{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\lambda},\boldsymbol{q},\boldsymbol{\rho},\boldsymbol{w}}{\text{minimize}} \quad \boldsymbol{c}^\top \boldsymbol{x}\\
&\text{subject to } \lambda_k \epsilon + \frac{1}{M}\sum_{i\in[M]} q_{ik} \leq 0 && \forall k \in [K]\\
&\qquad \delta(\boldsymbol{w}_{ink} + \sum_{n'\in[N]} \rho_{inkn'} \boldsymbol{a}_n^0 - \boldsymbol{a}_{n'}(\boldsymbol{x}) \mid \Xi) - \boldsymbol{w}_{ink}{}^\top \hat{\boldsymbol{\xi}}_i\\
&\qquad + \sum_{n'\in[N]} \rho_{inkn'} b_{n'}^0 - b_n(\boldsymbol{x}) - (\sum_{n'\in[N]} \rho_{inkn'} - c_n)\hat{t}_k - q_{ik} \leq 0 && \forall i \in [M], n \in [N+1], k \in [K]\\
&\qquad \|\boldsymbol{w}_{ink}\|_* \leq \lambda_k && \forall i \in [M], n \in [N+1], k \in [K]
\end{aligned}$$

$$\sum_{n'\in[N]} \rho_{inkn'} \leq 1 \qquad \forall i \in [M], n \in [N+1], k \in [K]$$

$$\lambda_k, \rho_{inkn'} \geq 0; q_{ik} \in \mathbb{R}; \boldsymbol{w}_{ink} \in \mathbb{R}^m \qquad \forall i \in [M], n \in [N+1], k \in [K], n' \in [N].$$

*Furthermore, it can be reformulated as a linear programming problem if $\mathcal{X}$ and $\Xi$ are polyhedral and Assumption 3 is satisfied.*

## 6. An Exact Solution Scheme

Inspired by Postek and den Hertog (2016) and Bertsimas and Dunning (2016), in this section we propose an exact solution scheme for the multistage robust optimization problem (8) by using an iterative partitioning method. In what follows, we first present the iterative partition based solution algorithm in Section 6.1, then describe how finite scenarios set $\hat{\mathcal{T}}$ gets updated in Section 6.2 and how to modify the partition $\mathscr{P}$ at each step in Section 6.3.

### 6.1. Iterative Partition based Solution Algorithm

We now show how the upper bound and lower bound can be updated iteratively. At the beginning, we initialize an original partition $\mathscr{P}^1 = \{\bar{\mathcal{T}}\}$ and an original scenario set $\hat{\mathcal{T}}^0 := \emptyset$. In each iteration $\ell = 1, 2, \ldots$, given the partition $\mathscr{P}^\ell$, we solve the upper bound problem (10) to obtain the optimal solutions $(\boldsymbol{x}^{*\ell}, \boldsymbol{\lambda}^{*\ell}, \boldsymbol{q}^{*\ell}, \bar{\boldsymbol{q}}^{*\ell})$ and the current upper bound $\mathrm{UB}^\ell$. By using the optimal solutions, we can generate a so-called "active scenario set" $\hat{\mathcal{A}}^\ell$, i.e., a set of scenarios considered to make constraints (10b) and (10c) binding. We provide the details on how to detect an active scenarios set in Section 6.2. We can then potentially improve the lower bound $\mathrm{LB}^\ell$ obtained from (11) by adding such active scenarios to $\hat{\mathcal{T}}^\ell$. We further attempt to improve the upper bound by exploiting the active scenarios to refine the partition $\mathscr{P}^{\ell+1} := \mathcal{V}(\mathscr{P}^\ell, \hat{\mathcal{A}}^\ell)$, which details can be found in Section 6.3. We repeat these steps until either the time or iteration limit is reached or when a sub-optimality of $\varepsilon$ has been confirmed. We present the pseudo-code of such an iterative partition based solution method in Algorithm 1.

The following proposition provides conditions under which one has monotonous improvement guarantees for the upper bound and lower bound generated by Algorithm 1.

Proposition 7. *When Algorithm 1 is followed, then $LB^{\ell+1} \geq LB^\ell$. Moreover, if for all $\mathcal{T}' \in \mathscr{P}^{\ell+1}$, there exists $\mathcal{T}'' \in \mathscr{P}^\ell$ such that $\mathcal{T}' \subseteq \mathcal{T}''$, then $UB^\ell \geq UB^{\ell+1}$.*

### 6.2. Detecting an Active Scenarios Set

In the following, we present a simple way to obtain an active scenarios set by identifying the scenarios of $t$, in which we expect that these active scenarios that bind the solutions with finitely adaptive policies are also binding the fully adaptive solutions.[1]

[1] A scenario $\hat{t}$ is considered "binding" in problem (10) either if constraint (10b) is active at $\hat{t}$ or if there exists a triplet $(\boldsymbol{\xi}, i, n)$ that makes constraint (10c) active at $\hat{t}$.

**Algorithm 1** Iterative Partition based Solution Algorithm

```
Initialize: $\text{LB}^0 = -\infty$, $\text{UB}^0 = +\infty$, $\bar{\mathcal{T}} := [\inf_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{\xi}), \sup_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{\xi})]$, $\mathscr{P}^1 := \{\bar{\mathcal{T}}\}, \ell = 1$, $\hat{\mathcal{T}}^0 := \emptyset$, $\varepsilon$.
while $|(\text{UB}^{\ell-1} - \text{LB}^{\ell-1})/\text{UB}^{\ell-1} \times 100\%| > \varepsilon$ do
    Solve the upper bound problem (10) with the partition $\mathscr{P}^\ell$.
    Identify the optimal solution $(\boldsymbol{x}^{*\ell}, \boldsymbol{\lambda}^{*\ell}, \boldsymbol{q}^{*\ell}, \bar{\boldsymbol{q}}^{*\ell})$ and optimal objective $\text{UB}^\ell$.
    Calculate an active scenarios set $\hat{\mathcal{A}}^\ell$.                          ▷ Section 6.2
    Update the finite scenarios set $\hat{\mathcal{T}}^\ell \leftarrow \hat{\mathcal{A}}^\ell \bigcup \hat{\mathcal{T}}^{\ell-1}$.
    Solve lower bound problem (11) with $\hat{\mathcal{T}}^\ell$ and identify the new lower bound $\text{LB}^\ell$.
    Update the partitions $\mathscr{P}^{\ell+1} \leftarrow \mathcal{V}(\mathscr{P}^\ell, \hat{\mathcal{A}}^\ell)$, and $l := l + 1$.      ▷ Section 6.3
end while
return optimal objective value $z^*$ and optimal solution $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{q}^*, \bar{\boldsymbol{q}}^*)$.
```

In order to detect the active scenarios at $\ell$-th iteration, let $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{q}^*, \bar{\boldsymbol{q}}^*)$ be the optimal solutions of problem (10) at the $\ell$-th iteration. For constraint (10b), one can identify an active scenario $\hat{t}$ as

$$\hat{t}_k^1 \in \arg\min_{t\in\mathcal{T}_k} \left\{ -\lambda_k^* \epsilon - \frac{1}{M} \sum_{i\in[M]} (\bar{q}_{ik}^* + q_{ik}^* t) \right\}. \tag{12}$$

Similarly, for constraint (10c), we can use a minimizer as

$$\hat{t}_k^2 \in \arg\min_{t\in\mathcal{T}_k} \left\{ \bar{q}_{ik}^* + q_{ik}^* t - g_{n'}(\boldsymbol{x}^*, \boldsymbol{\xi}, t) + \lambda_k^* \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i\|, \ \forall \boldsymbol{\xi} \in \Xi, i \in [M], n' \in [N+1] \right\}. \tag{13}$$

Finally, we let $\hat{\mathcal{A}} = \bigcup_{k\in[K]} \left\{ \hat{t}_k^1 \bigcup \hat{t}_k^2 \right\}$. Note that an alternative method for detecting the active scenarios is presented in Hadjiyiannis et al. (2011).

### 6.3. Updating the Partition

In the following, we describe how to update the partition $\mathscr{P}^{\ell+1}$ by exploiting a new set of active scenarios $\hat{\mathcal{A}}^\ell$ while satisfying the nested condition in Proposition 7. More specifically, given a partition $\mathscr{P}^\ell$, the new partition is constructed by using a Voronoi diagram partition:

$$\mathcal{V}\left(\mathscr{P}^\ell, \hat{\mathcal{A}}^\ell\right) := \bigcup_{\mathcal{T}\in\mathscr{P}^\ell} \bigcup_{\hat{t}\in\hat{\mathcal{A}}^\ell} \left( \mathcal{T} \cap \left\{ t \mid |\hat{t} - t| \le |\hat{t}' - t|, \ \forall \hat{t}' \in \hat{\mathcal{A}}^\ell \right\} \right).$$

The idea behind this operation is to create intervals that are "centered" at each active scenario.

In the following example, we illustrate how this operation works.

Example 1. Given a partition $\mathscr{P} := \{[0,3],[3,8],[8,10]\}$ and an active scenario set $\hat{\mathcal{A}} := \{1,6,7\}$, we obtain $\mathcal{V}\left(\mathscr{P}, \hat{\mathcal{A}}\right) = \{[0,3],[3,3.5],[3.5,6.5],[6.5,8],[8,10]\}$ when performing the above operation. Note that, for any members $\mathcal{T} \in \mathcal{V}\left(\mathscr{P}, \hat{\mathcal{A}}\right)$, there always exists a member $\mathcal{T}'$ in $\mathscr{P}$ for which $\mathcal{T} \subseteq \mathcal{T}'$.

## 7. Analysis of a Decomposable DRSSDCP Formulation

One might consider a number of different variations of the data-driven DRSSDCP presented in Section 5. The choice of the formulation might depend on the type of information that is at hand regarding the distributions of both the controlled and reference performance variables. For example, in some situations, the reference performance might be formulated based on a study of the past, while the controlled performance considers events that are to happen in the future. Alternatively, the information about the two variables could come from two separate datasets with no information about how the perturbations of one might affect the other, e.g. when comparing the distribution of lost sales for a new product compared to the distribution of an old one. This section focuses on a second variation of the DRSSDCP that can be used to handle these situations and describe how the methods presented in sections 5 and 6 might be adapted to it.

We start by considering that both the controlled performance function and reference performance function are now parameterized by $f(\boldsymbol{x},\boldsymbol{\xi})$ and $f_0(\boldsymbol{\zeta})$ respectively, with $\boldsymbol{\xi}\in\Xi_\xi\subset\mathbb{R}^m$ and $\boldsymbol{\zeta}\in\Xi_\zeta\subset\mathbb{R}^m$ denoting two possibly different sources of uncertainty. Let the new measurable space be the product space $(\Xi_\xi\times\Xi_\zeta,\mathcal{B}(\Xi_\xi\times\Xi_\zeta))$.

In this new environment, Assumption 1 can be reformulated as follows.

Assumption 4. *The feasible set $\mathcal{X}$ is a non-empty convex set, and the sets $\Xi_\xi$ and $\Xi_\zeta$ are non-empty compact convex.*

We are now interested in the case that the marginal distributions $\mathbb{P}_\xi$ and $\mathbb{P}_\zeta$ are only known to belong to their respective Wasserstein ambiguity sets. Hence, we will consider the following Decomposable-DRSSDCP:

$$[\text{D-DRSSDCP}]\quad \underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}}\ \boldsymbol{c}^\top\boldsymbol{x} \tag{14a}$$

$$\text{subject to}\ f(\boldsymbol{x},\boldsymbol{\xi})\succeq^{\mathbb{P}}_{(2)} f_0(\boldsymbol{\zeta}) \qquad \forall\mathbb{P}\in\mathcal{P}^r_{\mathrm{W}^2}, \tag{14b}$$

where $\mathcal{P}^r_{\mathrm{W}^2}:=\bigcup_{\mathbb{P}_\xi\in\mathcal{P}^r_\xi(\hat{\mathbb{P}}_\xi,\epsilon_\xi),\mathbb{P}_\zeta\in\mathcal{P}^r_\zeta(\hat{\mathbb{P}}_\zeta,\epsilon_\zeta)}\mathcal{M}(\mathbb{P}_\xi,\mathbb{P}_\zeta)$, and $\mathcal{M}(\mathbb{P}_\xi,\mathbb{P}_\zeta)$ is the set of all distributions of $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ with marginal distributions $\mathbb{P}_\xi$ and $\mathbb{P}_\zeta$, and where

$$\mathcal{P}^r_\xi(\hat{\mathbb{P}}_\xi,\epsilon_\xi):=\left\{\mathbb{P}_\xi\in\mathcal{M}(\Xi_\xi)\ \middle|\ d^r_{\mathrm{W}}(\mathbb{P}_\xi,\hat{\mathbb{P}}_\xi)\le\epsilon_\xi\right\},$$

and

$$\mathcal{P}^r_\zeta(\hat{\mathbb{P}}_\zeta,\epsilon_\zeta):=\left\{\mathbb{P}_\zeta\in\mathcal{M}(\Xi_\zeta)\ \middle|\ d^r_{\mathrm{W}}(\mathbb{P}_\zeta,\hat{\mathbb{P}}_\zeta)\le\epsilon_\zeta\right\}$$

with $\hat{\mathbb{P}}_\xi$ and $\hat{\mathbb{P}}_\zeta$ as the respective empirical distributions with $\{\hat{\xi}_i\}_{i=1}^{M_\xi}$ and $\{\hat{\zeta}_{i'}\}_{i'=1}^{M_\zeta}$.

We remark that this variation recovers the DRSSDCP (4) only when $\{\hat{\xi}_i\}_{i=1}^{M_\xi}=\{\hat{\zeta}_{i'}\}_{i'=1}^{M_\zeta}$, $\epsilon_\xi=\epsilon_\zeta$, and one additionally imposes that the marginals $\mathbb{P}_\xi=\mathbb{P}_\zeta$. The latter implies that D-DRSSDCP should not be considered a generalization of DRSSDCP but could serve as its conservative approximation. Furthermore, the ambiguity set could have alternatively been defined as $\mathcal{P}^r_{\mathrm{W}^2} := \bigcup_{\mathbb{P}_\xi \in \mathcal{P}^r_\xi(\hat{\mathbb{P}}_\xi, \epsilon_\xi), \mathbb{P}_\zeta \in \mathcal{P}^r_\zeta(\hat{\mathbb{P}}_\zeta, \epsilon_\zeta)} \mathbb{P}_\xi \times \mathbb{P}_\zeta$, given that from Lemma 1, we have that, for all $\mathbb{Q} \in \mathcal{M}(\mathbb{P}_\xi, \mathbb{P}_\zeta)$,

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{\xi}) \succeq^{\mathbb{Q}}_{(2)} f_0(\boldsymbol{\zeta}) \quad &\equiv \quad \forall \eta \in \mathbb{R},\ \mathbb{E}_{\mathbb{Q}}[(\eta - f(\boldsymbol{x}, \boldsymbol{\xi}))^+] \leq \mathbb{E}_{\mathbb{Q}}[(\eta - f_0(\boldsymbol{\zeta}))^+] \\ &\equiv \quad \forall \eta \in \mathbb{R},\ \mathbb{E}_{\mathbb{P}_\xi}[(\eta - f(\boldsymbol{x}, \boldsymbol{\xi}))^+] \leq \mathbb{E}_{\mathbb{P}_\zeta}[(\eta - f_0(\boldsymbol{\zeta}))^+] \\ &\equiv \quad \forall \eta \in \mathbb{R},\ \mathbb{E}_{\mathbb{P}_\xi \times \mathbb{P}_\zeta}[(\eta - f(\boldsymbol{x}, \boldsymbol{\xi}))^+] \leq \mathbb{E}_{\mathbb{P}_\xi \times \mathbb{P}_\zeta}[(\eta - f_0(\boldsymbol{\zeta}))^+] \\ &\equiv \quad f(\boldsymbol{x}, \boldsymbol{\xi}) \succeq^{\mathbb{P}_\xi \times \mathbb{P}_\zeta}_{(2)} f_0(\boldsymbol{\zeta}) \end{aligned}$$

In other words, constraint (14b) is insensitive to the type of correlation that are imposed between $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ in $\mathcal{P}^r_{\mathrm{W}^2}$.

We briefly discuss some special cases of D-DRSSDCP. First, it is straightforward to see that once again, D-DRSSDCP reduces to SDCP2 and DFSDCP (with constraint (5b) replaced with $f(\boldsymbol{x}, \boldsymbol{\xi}) \geq f_0(\boldsymbol{\zeta})$, $\forall (\boldsymbol{\xi}, \boldsymbol{\zeta}) \in \Xi_\xi \times \Xi_\zeta$) when $\epsilon_\xi = \epsilon_\zeta = 0$ and $\epsilon_\xi = \epsilon_\zeta = \infty$ respectively. The cases where either $\epsilon_\xi > \epsilon_\zeta = 0$ and $\epsilon_\zeta > \epsilon_\xi = 0$ allows one to model situations where the distribution of either the controlled performance or the reference performance is exactly known. The former is particularly useful in situations where we wish to describe the reference performance in terms of a distribution function rather than as a random variable. This is in particular the form that appears in Sehgal and Mehra (2020) as summarized in Appendix B.3.

Finally, the case where $\epsilon_\xi > 0$ and $\epsilon_\zeta > 0$ captures ambiguity about the distributions of both the controlled and reference variables with information possibly originating from two separate datasets. In the following proposition, we show that D-DRSSDCP (14) can be reformulated as a multistage robust optimization problem under mild conditions, which takes the form of $\min_{\boldsymbol{x}}$-$\sup_t$-$\min_{\lambda^1, \lambda^2, \boldsymbol{q}, \boldsymbol{r}}$-$\sup_{\boldsymbol{\xi}, \boldsymbol{\zeta}}$.

Proposition 8. *Under assumptions 2 and 4, D-DRSSDCP* (14) *coincides with the optimal value of the following multistage robust optimization problem:*

$$\begin{aligned} &\underset{\boldsymbol{x} \in \mathcal{X}}{\text{minimize}}\ \boldsymbol{c}^\top \boldsymbol{x} && && \text{(15a)} \\ &\text{subject to}\ H(\boldsymbol{x}, t) \leq 0 && \forall t \in \bar{\mathcal{T}}', && \text{(15b)} \end{aligned}$$

*where* $\bar{\mathcal{T}}' := [t'_{min}, t'_{max}]$, *with* $t'_{min} := \inf_{\boldsymbol{\zeta} \in \Xi_\zeta} f_0(\boldsymbol{\zeta})$ *and* $t'_{max} := \sup_{\boldsymbol{\zeta} \in \Xi_\zeta} f_0(\boldsymbol{\zeta})$, *and where*

$$\begin{aligned} H(\boldsymbol{x}, t) := \quad & \inf_{\lambda^1, \lambda^2, \boldsymbol{q}, \boldsymbol{r}} && \lambda^1 \epsilon_\xi + \lambda^2 \epsilon_\zeta + \frac{1}{M_\xi} \sum_{i \in [M_\xi]} q_i + \frac{1}{M_\zeta} \sum_{i' \in [M_\zeta]} r_{i'} && \text{(16a)} \\ & \text{subject to} && \sup_{\boldsymbol{\xi} \in \Xi_\xi} (t - f(\boldsymbol{x}, \boldsymbol{\xi}))^+ - \lambda^1 \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i\| \leq q_i \qquad \forall i \in [M_\xi] && \text{(16b)} \\ & && \sup_{\boldsymbol{\zeta} \in \Xi_\zeta} -(t - f_0(\boldsymbol{\zeta}))^+ - \lambda^2 \|\boldsymbol{\zeta} - \hat{\boldsymbol{\zeta}}_{i'}\| \leq r_{i'} \qquad \forall i' \in [M_\zeta] && \text{(16c)} \\ & && \lambda^1, \lambda^2 \geq 0; \boldsymbol{q} \in \mathbb{R}^{M_\xi}; \boldsymbol{r} \in \mathbb{R}^{M_\zeta}. && \text{(16d)} \end{aligned}$$

Following similar procedures to those discussed in Section 5, we can also obtain a conservative approximation of D-DRSSDCP by using finite adaptability, which provides an upper bound, and derive a lower bound by using a finite scenario set. For completeness, we provide the finite-dimensional reformulation of the conservative approximation for problem (15) in Appendix B.4. Moreover, by following similar steps, we can adapt Algorithm 1 to iteratively tighten the upper and lower bounds. For the sake of conciseness, we omit these details.

In the following, we will focus on the case where the reference performance function $f_0(\boldsymbol{\zeta})$ has a known distribution, i.e., $\epsilon_\zeta = 0$. In this case, $\mathcal{P}^1_\zeta(\hat{\mathbb{P}}_\zeta, \epsilon_\zeta)$ reduces to a singleton $\{\hat{\mathbb{P}}_\zeta\}$. Now we conclude this section with Proposition 9, which shows that D-DRSSDCP reduces to solving a finite-dimensional convex optimization problem.

Proposition 9. *Given that $\epsilon_\xi > \epsilon_\zeta = 0$, and that $\mathcal{P}^1_{W^2}$ is used, under assumptions 2 and 4, D-DRSSDCP* (14) *reduces to the following finite-dimensional convex optimization problem,*

$$
\begin{aligned}
&\underset{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\lambda},\boldsymbol{s},\boldsymbol{v}}{\text{minimize}} \ \boldsymbol{c}^\top \boldsymbol{x} && && \text{(17a)}\\
&\text{subject to } \lambda_j \epsilon_\xi + \frac{1}{M_\xi}\sum_{i\in[M_\xi]} s_{ij} \le \gamma_j && \forall j \in [M_\zeta] && \text{(17b)}\\
&\qquad \delta(\boldsymbol{v}_{ijn}|\Xi_\xi) - \boldsymbol{w}_{ijn}^\top \hat{\boldsymbol{\xi}}_i - b_n(\boldsymbol{x}) + c_n t_j \le s_{ij} && \forall i \in [M_\xi], j \in [M_\zeta], n \in [N+1] && \text{(17c)}\\
&\qquad \|\boldsymbol{w}_{ijn}\|_* \le \lambda_j && \forall i \in [M_\xi], j \in [M_\zeta], n \in [N+1] && \text{(17d)}\\
&\qquad \boldsymbol{w}_{ijn} = \boldsymbol{v}_{ijn} + \boldsymbol{a}_n(\boldsymbol{x}) && \forall i \in [M_\xi], j \in [M_\zeta], n \in [N+1] && \text{(17e)}\\
&\qquad \lambda_j \ge 0, \boldsymbol{v}_{ijn} \in \mathbb{R}^m, s_{ij} \in \mathbb{R} && \forall i \in [M_\xi], j \in [M_\zeta], n \in [N+1], && \text{(17f)}
\end{aligned}
$$

*where $a_{N+1} = b_{N+1} = c_{N+1} = 0$ and $c_n = 1$ for all $n \in [N]$, $\gamma_j = \frac{1}{M_\zeta}\sum_{i\in[M_\zeta]} \Big(t_j - f_0(\hat{\boldsymbol{\zeta}}_i)\Big)^+$ and $t_j = f_0(\hat{\boldsymbol{\zeta}}_j)$ for all $j \in [M_\zeta]$. Moreover, it can be reformulated as a linear programming problem if $\mathcal{X}$ and $\Xi_\xi$ are polyhedral and Assumption 3 is satisfied.*

## 8. Numerical Study

In this section, we consider two applications to illustrate our DRSSDCP modeling paradigm and proposed solution scheme. For the sake of brevity, we focus on the DRSSDCP discussed in Section 5. We will investigate the computational efficiency of our iterative partitioning algorithm and the effect of changing the radius of the Wasserstein ball on out-of-sample performance in the contexts involving both synthetic and real-world data. Section 8.1 considers a simple resource allocation problem where the marginal revenue of projects is modeled using independent continuous distributions. Here, we expect that data-driven robustification (using DRSSDCP with the Wasserstein ambiguity set) will allow us to identify better performing solutions, compared to those produced using a SAA scheme, when evaluated on the true underlying distribution. Section 8.2 considers a more realistic portfolio optimization problem where historical observations are used to approximate the future distribution of stock returns. We are interested in verifying whether a DRSSDCP can effectively mitigate the fact that real stock returns processes do not satisfy the i.i.d. assumption made by empirical risk models.

On the technical side, in both applications, we will use the $\ell_1$-norm as the reference distance metric, namely, $d(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) := \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|_1$. Also, all our experiments are implemented in C programming language and use IBM CPLEX solver, version 12.10.0 callable libraries with the default settings. All experiments are conducted on the cedar cluster of Compute Canada. When solving a DRSSDCP, the algorithm proposed in Section 6 is run until either an optimality gap of 1% or a maximum of 2 hours time limit is reached. If the instances cannot be solved optimally within the time limit, we choose the most recent solutions of the conservative approximation model as the optimal solution.

Remark 3. DRSSD constraints are known to suffer from numerical issues because they do not satisfy Slater's constraint qualification (e.g., Hu et al. 2011, Guo et al. 2017, Hu et al. 2012, Chen and Jiang 2018), for this reason, in the following two numerical studies, we will consider the slightly relaxed $\phi$-DRSSDCP where constraint (6b) takes the form:

$$\mathbb{E}_{\mathbb{P}}\left[(t - \boldsymbol{\xi}^\top \boldsymbol{x})^+\right] \leq \mathbb{E}_{\mathbb{P}}\left[(t - \boldsymbol{\xi}^\top \boldsymbol{x}_0)^+\right] + \phi, \, \forall t \in \mathbb{R}, \, \forall \mathbb{P} \in \mathcal{P}_{\mathrm{W}}^1(\hat{\mathbb{P}}, \epsilon)$$

with $\phi = 0.01$.

### 8.1. A Simple Resource Allocation Problem

In this section, we conduct our numerical experiments on a simple resource allocation problem with DRSSD constraints. More specifically, we are interested in determining the optimal allocation $\boldsymbol{x} := [x_1, x_2, x_3]^\top$ of the total resources to invest in a set of 3 projects, with $x_1 + x_2 + x_3 = 1$ and $x_j \geq 0, \forall j \in [3]$. The marginal revenue of each project is denoted by $\boldsymbol{\xi} := [\xi_1, \xi_2, \xi_3]^\top$, and considered uncertain. The information available about $\boldsymbol{\xi}$ consists in a set of i.i.d. observations $\left\{\hat{\boldsymbol{\xi}}^1, \hat{\boldsymbol{\xi}}^2, \cdots, \hat{\boldsymbol{\xi}}^M\right\}$. This motivates the use of a Wasserstein set $\mathcal{P}_{\mathrm{W}}^1(\hat{\mathbb{P}}, \epsilon)$, with $\hat{\mathbb{P}}$ as the empirical distribution and a box support set $\Xi := [0, \, 10]^3$ assumed to contain the support of the true underlying distribution $\bar{\mathbb{P}}$.[2] A natural DRSSDCP that can be used in this context lets $c := \mathbb{E}_{\hat{\mathbb{P}}}[\boldsymbol{\xi}]$, $f(\boldsymbol{x}, \boldsymbol{\xi}) := \boldsymbol{\xi}^\top \boldsymbol{x}$, and $f_0(\boldsymbol{\xi}) := \boldsymbol{\xi}^\top \boldsymbol{x}_0$, where $\boldsymbol{x}_0 = [1, 0, 0]^\top$ is a reference strategy that invests all resources in project #1. In words, the objective is to maximize the expected revenue based on the empirical distribution while enforcing that the revenue of the selected projects robustly stochastically dominates the revenue of a reference allocation strategy. Note that we omit to include a worst-case expected revenue objective function in order to focus our attention on statistical robustness of the SSD constraint.

[2] In our experiments, this box set ended up covering 99.8% of the mass of the unknown measure.

In the following, we first describe the details of the synthetic data instances in Section 8.1.1, then show the computational performance of our proposed algorithm in terms of different $\epsilon$ and $M$ in Section 8.1.2. We discuss some numerical results regarding the average optimal allocation for different in-sample sizes $M$ as a function of $\epsilon$ in Section 8.1.3. Finally, we present out-of-sample performance of DRSSCP solutions with respect to different values of $\epsilon$ and $M$ in Section 8.1.4.

**8.1.1. Synthetic Data Generation Scheme** Our numerical experiments employ synthetic data in order to create an environment in which the observations that are used by our data-driven model are drawn from the same distribution as those used to measure the out-of-sample performance. Specifically, we consider that the underlying distribution $\bar{\mathbb{P}}$ is designed such that the three projects are independent from each other, while their respective marginal distribution is such that they satisfy $\xi_3 \succ_{(2)}^{\bar{\mathbb{P}}} \xi_1 \succ_{(2)}^{\bar{\mathbb{P}}} \xi_2$ and $\mathbb{E}_{\bar{\mathbb{P}}}[\xi_3] > \mathbb{E}_{\bar{\mathbb{P}}}[\xi_2] = \mathbb{E}_{\bar{\mathbb{P}}}[\xi_1]$. This means that theoretically it is optimal to invest all the resources in project #3 since it stochastically dominates $\xi_1$ and achieves the highest expected revenue. On the other hand, it is theoretically infeasible to invest all resources in project #2. To be precise about $\bar{\mathbb{P}}$, we first define the distribution of the marginal revenue of project #1, i.e. $\xi_1$, using an auxiliary random variable $\boldsymbol{z}_0 \sim \mathrm{logN}(0.08, 0.03)$, i.e. with a lognormal distribution such that $\mathbb{E}[\boldsymbol{z}_0] = 0.08$ and $(\mathbb{E}[(\boldsymbol{z}_0 - 0.08)_2])^{1/2} = 0.03$, to get $\xi_1 \sim \mathrm{logN}(\boldsymbol{z}_0, 5)$. We then have that $\xi_2$ is independently distributed as $\xi_2 \sim \mathrm{logN}({\xi_1}', 5)$, where ${\xi_1}'$ is i.i.d. to $\xi_1$. Finally, we have that $\xi_3 = 1.125{\boldsymbol{z}_0}'$ where ${\boldsymbol{z}_0}'$ is i.i.d. to $\boldsymbol{z}_0$. [3]

Each experiment will involve $M \in \{10, 100, 1000\}$ empirical samples that are identically and independently generated from $\bar{\mathbb{P}}$. For each size $M$, we repeat 100 experiments in which the out-of-sample performance is measured using a second set of 10,000 independent samples.

**8.1.2. Computational Performance** Table 1 reports on the numerical efficiency of our iterative partition based solution algorithm. One can first observe that, while most instances (i.e., 99.4%) are solved in less than 2 hours, the solution time increases as $M$ increases. Perhaps more interestingly, it appears that some computational difficulties appear for midrange values of $\epsilon$, which causes more rounds of partitions. The reported numerical efficiency for large $\epsilon$ might be due to the fact that the DRSSDCP reduces to the DFSDCP which is insensitive to the size of $\mathcal{T}$ and $M$.

[3] Indeed, we have that, for any given $t \geq 0$, $\mathbb{P}(\xi_3 \geq t) = \mathbb{P}(1.125{\boldsymbol{z}_0}' \geq t) = \mathbb{P}({\boldsymbol{z}_0}' \geq t/1.125) \geq \mathbb{P}({\boldsymbol{z}_0}' \geq t)$. Based on the definition of FSD, we have $\xi_3 \succeq_{(1)} {\boldsymbol{z}_0}'$. Given that ${\boldsymbol{z}_0}'$ is i.i.d. to $\boldsymbol{z}_0$, so $\xi_3 \succeq_{(1)} \boldsymbol{z}_0$. Since FSD implies SSD, we further have that $\xi_3 \succeq_{(2)} \boldsymbol{z}_0$. For any given $t$, we also have $\mathbb{E}[(\xi_1 - t)^+] = \mathbb{E}[\mathbb{E}[(\xi_1 - t)^+ | \boldsymbol{z}_0]] \geq \mathbb{E}[(\mathbb{E}[\xi_1 | \boldsymbol{z}_0] - t)^+] = \mathbb{E}[(\boldsymbol{z}_0 - t)^+]$, where the inequality comes from the Jensen's inequality. Moreover, the inequality is strict when $t = \mathbb{E}[z_0]$. Based on the definition of SSD, we say $\boldsymbol{z}_0 \succ_{(2)} \xi_1$. Therefore, $\xi_3 \succ_{(2)} \xi_1$. On the other hand, we have $\mathbb{E}[(\xi_2 - t)^+] = \mathbb{E}[\mathbb{E}[(\xi_2 - t)^+ | {\xi_1}']] \geq \mathbb{E}[(\mathbb{E}[\xi_2 | {\xi_1}'] - t)^+] = \mathbb{E}[({\xi_1}' - t)^+]$, with a strict inequality when $t = \mathbb{E}[\xi_1]$, so ${\xi_1}' \succ_{(2)} \xi_2$. Since ${\xi_1}'$ is i.i.d. to $\xi_1$, then $\xi_1 \succ_{(2)} \xi_2$. We thus conclude that $\xi_3 \succ_{(2)} \xi_1 \succ_{(2)} \xi_2$. The relation between the expected revenues is more straightforward.

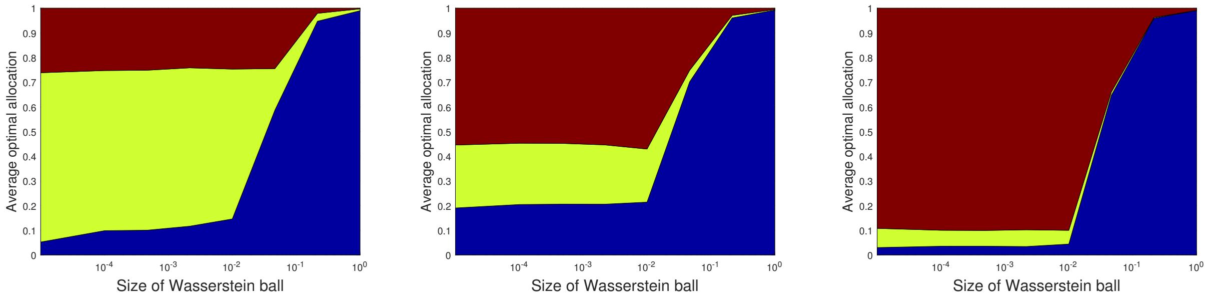
**Table 1** The average computational performance with respect to different $\epsilon$ and in-sample sizes ($M \in \{10, 100, 1000\}$), in terms of average CPU time (Time, in seconds), proportion of unsolved instances (prop, in %) within the 2 hour limit, and average number of iterations (# of Iter).

| $M$ | 10 | | | 100 | | | 1000 | | |
|---|---|---|---|---|---|---|---|---|---|
| $\epsilon$ | Time | prop | # of Iter | Time | prop | # of Iter | Time | prop | # of Iter |
| 0.0001 | 7.7 | 0.0 | 2.5 | 14.6 | 0.0 | 2.7 | 98.8 | 0.0 | 3.1 |
| 0.0005 | 7.2 | 0.0 | 2.7 | 15.8 | 0.0 | 2.7 | 108.6 | 0.0 | 3.1 |
| 0.0022 | 7.8 | 0.0 | 3.2 | 17.8 | 0.0 | 3.0 | 142.7 | 0.0 | 3.4 |
| 0.01 | 152.8 | 0.0 | 7.3 | 148.4 | 0.0 | 5.1 | 624.2 | 0.0 | 5.0 |
| 0.0464 | 52.7 | 0.0 | 11.1 | 466.6 | 0.0 | 7.1 | 5393.3 | 0.12[10.8] | 7.8 |
| 0.2154 | 6.8 | 0.01[1.75] | 9.6 | 15.4 | 0.0 | 3.6 | 44.8 | 0.0 | 1.5 |
| 1 | 2.6 | 0.0 | 5.5 | 3.4 | 0.0 | 1.1 | 12.1 | 0.0 | 1.0 |
| Average | 34.0 | - | 5.9 | 97.4 | - | 3.6 | 917.8 | - | 3.6 |

[ · ] in column of prop reports the average sub-optimality gap (in %) for the unsolved instances within the time limit.

**8.1.3. Analysis of Optimal Allocation** Figure 1 visualizes the impact of the size of the Wasserstein ball $\epsilon$ on the average optimal allocation policies for $M \in \{10, 100, 1000\}$ over 100 runs. As we can see from the figure, for all empirical sample sizes, the average optimal allocation gradually tends to turn into the reference strategy as $\epsilon$ increases. This can be explained by the fact that the feasible set shrinks as the ambiguity set becomes larger while the reference strategy always remains feasible since $f(\boldsymbol{x}_0, \boldsymbol{\xi}) \succeq_{(2)}^{\mathbb{P}} f_0(\boldsymbol{\xi})$ for all $\mathbb{P}$. More interestingly, we can also observe that when $M$ is small (see Figure 1(a)) the SDCP2 model (captured by the case where $\epsilon \to 0$) ends up recommending on average a large investment of 68% in project #2, which by construction is strictly stochastically dominated by the reference project #1. This is clear evidence of the optimizer's curse which is also referred as the problem of overfitting the small set of observations. Luckily, this issue appears to be partially resolved by using the DRSSDCP after properly sizing the ambiguity set. Namely for smaller $M = 10$, a larger $\epsilon$ can be used to recover a portfolio that is more similar to the reference one, while for larger $M = 1000$ a smaller $\epsilon$ will mostly suggest investing in the optimal project #3. This will further be confirmed in the following out-of-sample analysis. Alternatively, one can also confirm in Figure 1(c) that with a small radius, the DRSSDCP recovers the SAA solution, which is an allocation that is nearly optimal (i.e., fully invested in project #3) when the sample size is large enough.

**8.1.4. Out-of-Sample Performance of DRSSDCP solutions** We now turn to evaluating the out-of-sample performance of the different policies generated by the DRSSDCP as we change the size of the Wasserstein ball $\epsilon$. In particular, one might start by considering the out-of-sample expected revenue which statistics (over 100 experiments) are presented in Figure 2. As we can see



(a) $M = 10$ (b) $M = 100$ (c) $M = 1000$

**Figure 1** The average optimal allocation as a function of the size of Wasserstein ball $\epsilon$ for (a) $M = 10$, (b) $M = 100$, (c) $M = 1000$ training samples over 100 runs. The blue, yellow and red regions represent the average optimal allocation of resources for the project #1, project #2 and project #3, respectively.

from the figure, for $M \in \{10, 100, 1000\}$ the average out-of-sample expected revenue decreases as $\epsilon$ increases yet always outperforms the average expected revenue (dashed line) that is achieved by the reference strategy. This improved performance is gradually lost as we increase $\epsilon$, being sacrificed to improve SSD feasibility. The performance also improves as $M$ increases given that the empirical distribution model becomes a better representation of the underlying theoretical model.

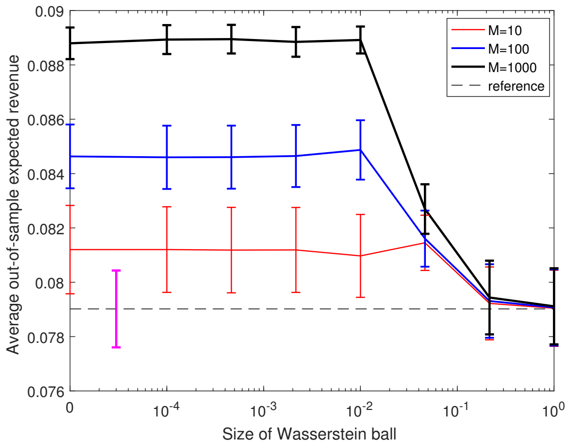
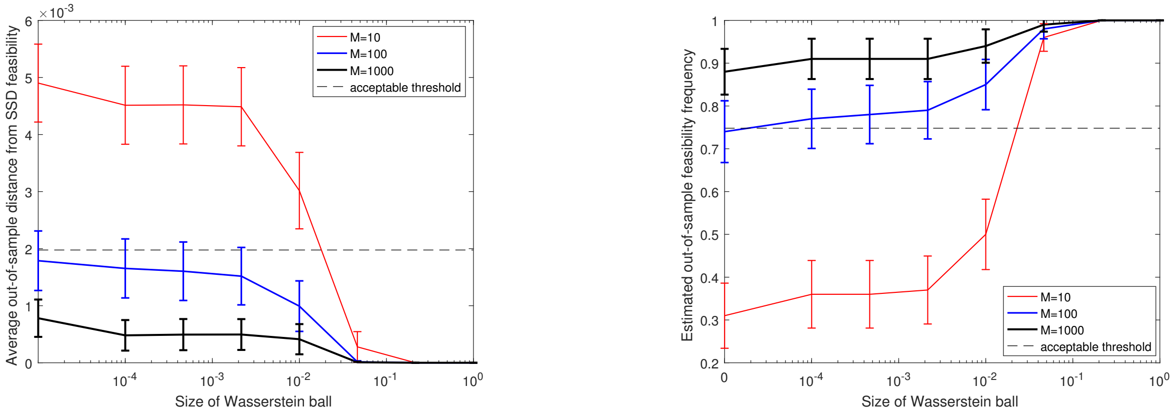


**Figure 2** Statistics of the out-of-sample expected revenue as a function of $\epsilon$ (each with 10,000 testing samples) for $M \in \{10, 100, 1000\}$. Solid lines indicate the average while the confidence bars identify the 10-th and 90-th percentiles based on the 100 runs. Finally, the dashed line and pink confidence bar show the statistics of the expected revenue achieved by the reference strategy.

In order to explore out-of-sample SSD feasibility, we introduce the notion of *out-of-sample distance from SSD feasibility*, which measures how far the proposed DRSSD policy is from being SSD feasible with respect to the out-of-sample distribution. Appendix B.5 provides further details about this measure which can be summarized as the type-1 Wasserstein distance of the out-of-sample distribution to its projection on the set of out-of-sample feasible distributions. Alternatively, we will also be interested in estimating the *out-of-sample feasibility frequency*, i.e. the probability of obtaining a DRSSDCP solution that satisfies the SSD constraint out-of-sample. Figure 3 presents statistics of both of these measures for $M \in \{10, 100, 1000\}$, as $\epsilon$ varies. We can observe from this figure that the out-of-sample feasibility is improved as $\epsilon$ increases, which is inline with our previous observation that the optimal allocation converges to the reference strategy as $\epsilon$ varies from 0 to 1.



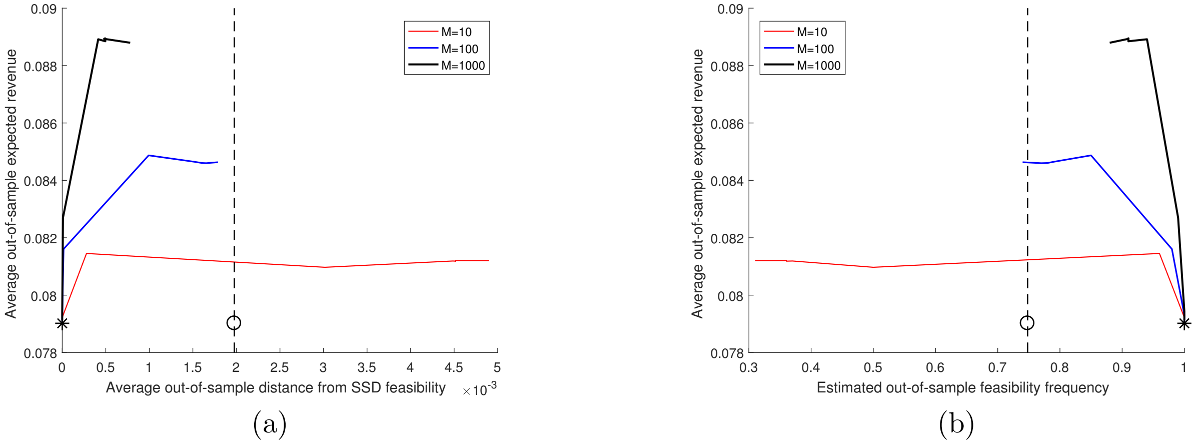
(a) (b)

**Figure 3** Statistics of the out-of-sample feasibility. (a) presents the mean (solid), 10-th, and 90-th percentiles (bars) of the out-of-sample distance from SSD feasibility. (b) presents the estimated out-of-sample feasibility frequency (with 90% confidence intervals) as a function of $\epsilon$ for $M \in \{10, 100, 1000\}$ empirical samples. The dashed lines identify an acceptable level of performance that is based on the mean out-of-sample distance from SSD feasibility (in (a)) and the out-of-sample feasibility estimate (in (b)) of a re-sampled version of the reference strategy's revenue distribution.

Interestingly, even when $M$ is large, one sees that one still needs a large amount of robustness in order to achieve a nearly perfect level of feasibility (say a feasibility frequency above 95%). Unfortunately, the experiments reveal that near perfect out-of-sample feasibility is only achieved by using a strategy that is nearly identical to the reference one. This is in contradiction with the fact that by construction project #3 is theoretically SSD feasible and achieves higher expected return. This leads us to conclude that near perfect out-of-sample feasibility is too strict of a criterion to aspire to. Alternatively, we consider a hypothetical project (call it project #1') which revenue is identically and independently distributed to the revenue of the reference project #1 and will use its out-of-sample feasibility performance as a threshold to identify strategies that have an "acceptable" level of out-of-sample feasibility. In simple words, a strategy will be considered out-of-sample acceptable if its out-of-sample performance is at least as feasible as a strategy that has the same revenue distribution as the reference strategy while being independent from it. This so-called "acceptable threshold" is presented in Figure 3 as the dashed lines. Note that based on the asymptotic convergence of empirical distributions, this acceptable level of feasibility is expected to converge to imposing nearly perfect SSD feasibility as $M$ goes to infinity.

Studying more closely which strategy is out-of-sample acceptable, we notice that when the in-sample size is small (i.e., $M = 10$), the average out-of-sample feasibility frequency fails to surpass the acceptable threshold if $\epsilon$ is also small (i.e., $\epsilon \leq 0.04$). This can however be fixed either by increasing the in-sample size (i.e. at $M = 100$ or $M = 1000$) or by increasing the radius of the Wasserstein ball (i.e. for $\epsilon > 0.04$). This confirms that DRSSDCP is an effective modeling paradigm to employ in a data-driven context where the number of observations is fixed.

Finally, Figure 4 presents two sets of bi-objective performance curves based on the strategies produced by the DRSSDCP as $\epsilon$ varies. We can clearly see that for any fixed in-sample size, it is possible to calibrate $\epsilon$ to identify strategies that outperform (by up to 12.5%) the reference strategy in terms of out-of-sample expected revenue while achieving an acceptable level of SSD feasibility.



(a)

(b)

**Figure 4** Bi-objective out-of-sample performance curves for strategies produced by the DRSSDCP. (a) presents the average out-of-sample expected revenue vs. the average out-of-sample distance from SSD feasibility. (b) presents resents the average out-of-sample expected revenue vs. the estimated out-of-sample feasibility frequency. In both figures, the star and the circle indicate respectively the performance achieved by the reference strategy and a re-sampled version of the reference strategy's revenue distribution.

### 8.2. A Data-driven Portfolio Optimization using Stock Market Data

We now turn to experimenting with the DRSSDCP in a real data-driven environment. Namely, we consider a portfolio optimization problem with an SSD constraint. Specifically, we are interested in choosing the proportions $\boldsymbol{x} \in \mathbb{R}^m$ of our total wealth to invest in each of $m$ assets, while assuming for simplicity that short-selling is forbidden. This leads to a similar definition of $\mathcal{X}$ as in Section 8.1, i.e., $\sum_{j=1}^{m} x_j = 1$ and $\boldsymbol{x} \geq 0$. The vector of weekly random return of each asset is denoted by $\boldsymbol{\xi} \in \mathbb{R}^m$. We will consider that the information available about the distribution of $\boldsymbol{\xi}$ consists in a set of historical observations of the $M$ most recent weekly returns, $\left\{\hat{\boldsymbol{\xi}}^1, \hat{\boldsymbol{\xi}}^2, \cdots, \hat{\boldsymbol{\xi}}^M\right\}$. This motivates the use of a Wasserstein ambiguity set $\mathcal{P}_{\mathrm{W}}^1(\hat{\mathbb{P}}, \epsilon)$, with $\hat{\mathbb{P}}$ as the empirical distribution and a box support set $\Xi := \{\boldsymbol{\xi} \in \mathbb{R}^m \mid \boldsymbol{\xi}^- \leq \boldsymbol{\xi} \leq \boldsymbol{\xi}^+\}$ assumed to contain the support of a "true" underlying distribution $\bar{\mathbb{P}}$.[4] Since the portfolio optimization problem is also a class of resource allocation problem, similarly, let $c := \mathbb{E}_{\hat{\mathbb{P}}}[\boldsymbol{\xi}]$, $f(\boldsymbol{x}, \boldsymbol{\xi}) := \boldsymbol{\xi}^\top \boldsymbol{x}$, and $f_0(\boldsymbol{\xi}) := \boldsymbol{\xi}^\top \boldsymbol{x}_0$, where $\boldsymbol{x}_0 := [1/m, 1/m, \ldots, 1/m]^\top$ is a reference portfolio that invests uniformly in all assets. The objective is to maximize the empirical expected return while enforcing that the return of the selected portfolio robustly stochastically dominates the returns of the reference portfolio.

The main purpose of this case study is to further explore the robustness of DRSSDCP solutions in a real data-driven environment, i.e. using stock market data, where both in-sample and out-of-sample realizations might not satisfy the i.i.d. assumption. Specifically, in the following we first describe our stock market data in Section 8.2.1. In Section 8.2.2, we show how the computational performance of our algorithm is affected by the number of stocks and samples. We then present a cross-validation scheme for selecting $\epsilon$ and $M$ in Section 8.2.3. Finally, we present the out-of-sample performance of our DRSSDCP solutions in Section 8.2.4.

**8.2.1. Stock Market Data Description** Simillarly to the work of Delage et al. (2021) and Delage and Li (2018), our case study uses the stock market data of the weekly returns of the stock value of companies that compose the S&P 500 index during the period spanning from January 1994 to December 2019. We partition the data into two parts, where the first part (called *in-sample* data) is used for model selection (i.e., $M$ and $\epsilon$ in problem (6)), while the second part (called *out-of-sample* data) is used for comparing the out-of-sample performance of our calibrated DRSSDCP model to an SAA approach. The *in-sample* data spans the period from January 1994 to December 2013 and include the 335 companies of the S&P 500 index that were continuously part of the index during this period. The *out-of-sample* data covers the next period from January 2014 to December 2019 and similarly the 257 companies that were present through that whole period. We note that the in-sample data is also used in our study of the computational performance of Algorithm 1.

**8.2.2. Computational Performance** In this section, we further show the computational performance for the more realistic-sized problems. We consider a set of stocks of size $m \in \{10, 50, 100\}$ and a set of observations of size $M \in \{50, 100\}$ and randomly generate 20 instances for each pair of $m$-$M$ using the in-sample data. We also focus on $\epsilon \in \{0.01, 0.0464\}$ as these midrange

[4] The values of $\xi_j^-$ and $\xi_j^+$ were chosen in a way that the box covers all realizations observed in the last 4 years.

**Table 2** The average computational performance with respect to different $\epsilon$, number of stocks ($m \in \{10, 50, 100\}$) and in-sample sizes ($M \in \{50, 100\}$), in terms of average CPU time (Time, in seconds), proportion of unsolved instances (prop, in %), and average number of iterations (# of Iter) over 20 runs.

| $m$ | | 10 | | | 50 | | | 100 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $M$ | $\epsilon$ | Time | prop | # of Iter | Time | prop | # of Iter | Time | prop | # of Iter |
| 50 | 0.01 | 242.3 | 0 | 6.0 | 1997.1 | 0.05[1.90] | 5.0 | 3979.3 | 0.25[2.25] | 5.0 |
| | 0.0464 | 103.2 | 0 | 5.0 | 3418.1 | 0.10[1.47] | 6.0 | 5966.3 | 0.25[3.83] | 6.0 |
| 100 | 0.01 | 1528.0 | 0 | 6.0 | 6259.1 | 0.05[3.01] | 6.0 | 6269.0 | 0.45[4.59] | 5.0 |
| | 0.0464 | 506.7 | 0 | 5.0 | 5966.3 | 0.25[3.83] | 6.0 | 5073.3 | 0.8[2.21] | 6.0 |
| Average | | 595.1 | 0 | 5.5 | 4410.2 | 0.11[2.55] | 5.8 | 5322.0 | 0.44[3.22] | 5.3 |

[ · ] in column of prop reports the average sub-optimality gap (in %) for the unsolved instances within 2 hours limit.

values appeared to be the hardest to handle in Section 8.1.2 (see Table 1). All the measurements are averages based on 20 runs.

Similarly to Table 1, Table 2 reports the numerical efficiency of our iterative partition based solution algorithm for a realistic-sized number of stocks and empirical samples. As we can see from the table, our algorithm can solve nearly 82% of the "hardest" problem instances optimally within 2 hours limit. The average solution time is increasing as the number of stocks $m$ and size of observation set $M$ increase. However, we remark that the average sub-optimality gap for instances that were not solved optimally within 2 hours limit, is relatively small, i.e., close to 2.55% and 3.22% on average for 50 and 100 stocks respectively. We suspect that most of these harder instances should be solved to a 1% gap in less than 5 hours.

**8.2.3. A Cross-validation Scheme for Selecting $\epsilon$ and $M$** This section employs the in-sample data set to calibrate the $\epsilon$ and $M$ parameters of the DRSSDCP in order to account for the non-stationarities that are present in our stock market data when composing a portfolio of $m = 5$ assets drawn randomly from this market. Our approach is inspired by sliding window cross-validation methods for time-series, which would suggest to create blocks of training and validation data by progressively passing a window through the whole in-sample period for each set of $m = 5$ assets. To reduce computing time, we instead create such blocks by first randomly selecting $m = 5$ companies from the pool of 335 and randomly pick a time point in the 19-years period. The sampled training block then consists of the 208 most recent weekly returns at that date for the 5 companies, while the validation block consists of the next $M' = 26$ weeks of returns. Model selection will therefore be based on the performance, measured with respect to the empirical distribution of the validation block, of the solution of a DRSSDCP that employs the most recent $M$ weekly return observations to define $\hat{\mathbb{P}}$. Our experiment involves 1000 such runs and our results report different statistics of the expected return, distance from SSD, and feasibility frequency over these runs.

We consider $M$ as ranging among $\{12, 52, 104, 208\}$ to represent quarterly, yearly, two-year and four-year periods of historical observations, while $\epsilon$ will span the range from 0 to 1.

The performance regarding expected return, distance from SSD, and feasibility frequency of the different model configurations are presented respectively in figures 5 and 6. Finally, Figure 7 presents the bi-objective curves which allow us to identify for each $M$ the optimal level of robustness that needs to be applied in order to maximize average expected returns in the validation runs while ensuring that the portfolios produced by DRSSDCP have an acceptable level of feasibility.

As in Section 8.1.4, it is clear that once again increasing robustness ($\epsilon$) leads to better feasibility on the validation data at the price of a reduction in expected return. One should remark however that the effect of historical sample size is different in this real-world setting. Indeed, it is not the case anymore that a larger set of historical observations necessarily leads to an improvement in out-of-sample performance. For instance, when $\epsilon$ is small, both the expected return and SSD feasibility is maximized by using $M = 52$. We believe this is due to the fact that stock return processes are non-stationary and therefore that optimization models that are based on empirical distributions suffer from higher generalization errors as $M$ becomes too large (i.e., $M > 52$).

Regarding the notion of acceptable feasibility, we note that since, in this real-world setting, we do not have access to the “true” underlying distribution, a different procedure therefore needs to be followed in order to produce an acceptable SSD feasibility thresholds. In the spirit of bootstrapping methods, a natural approach consists in comparing the feasibility measurement to the performance achieved by an investment that has its “true” underlying distribution exactly identical (and independent) to the realized empirical distribution of the reference portfolio in the validation block. To be more specific, in the case of Figure 6(a), the acceptable level of feasibility (i.e. the dashed line) consists in the average distance to SSD feasibility of the empirical distribution obtained by re-sampling (with replacement) $M'$ return scenarios from the returns produced with the reference portfolio in the validation block.

Given this new definition of acceptable feasibility, we conclude based on studying figures 5, 6 and 7 that the optimal choice of hyper-parameters in this case study consists in $M \in \{52, 104, 208\}$ and $\epsilon = 0.01$. Indeed, under these settings, the portfolio becomes acceptable with respect to the feasibility on the validation data while having a significantly better performance in terms of expected return compared to all other acceptable ones. In particular, the expected returns is nearly 0.029 p.p. higher than with the reference portfolio, and at least 0.02 p.p. higher than with the best acceptable portfolio achieved when $M = 12$. We suspect this is respectively due to the fact that the uniform portfolio does not take into account any historical information about potential future expected returns, and to the fact that when $M = 12$, the historical information is too noisy and causes large generalization errors. If asked to choose among $M \in \{52, 104, 208\}$, we would recommend $M = 52$

given that, as mentioned above, it appears that for low level of robustification (i.e., $\epsilon < 0.01$), the average performance under $M = 52$ dominates what is achieved under $M = 104$ and $M = 208$.

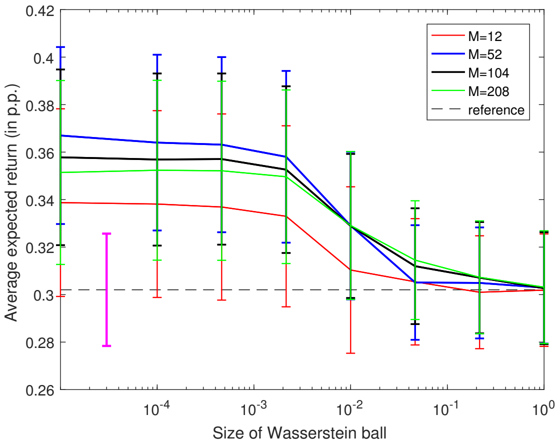


**Figure 5** Statistics of the expected return (in p.p.) on validation data as a function of $\epsilon$ over a time period of 26 weeks for $M \in \{12, 52, 104, 208\}$. Solid lines indicate the average while the confidence bars identify the 10-th and 90-th percentiles based on the 1000 runs. Finally, the dashed line and pink confidence bar show the statistics of the expected return achieved by the reference portfolio.

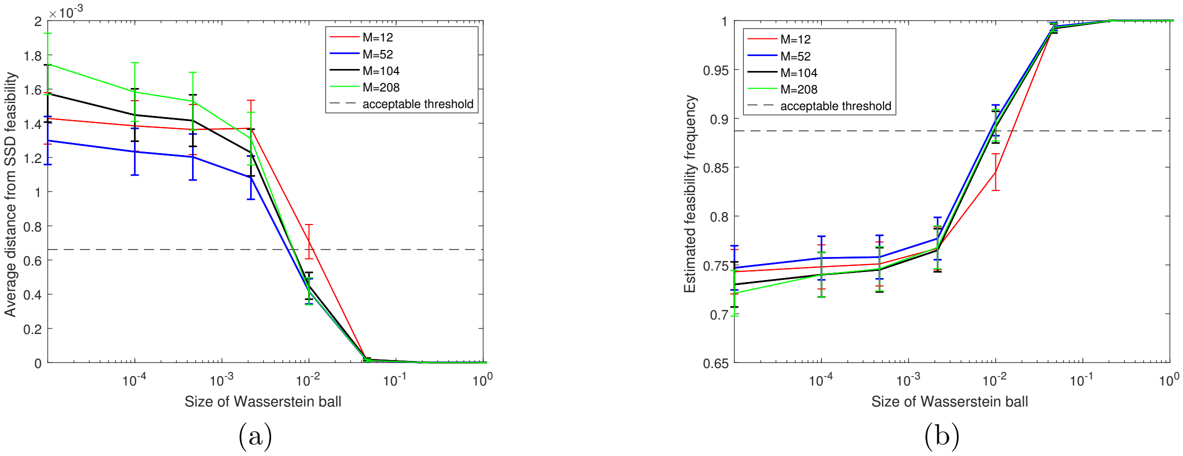
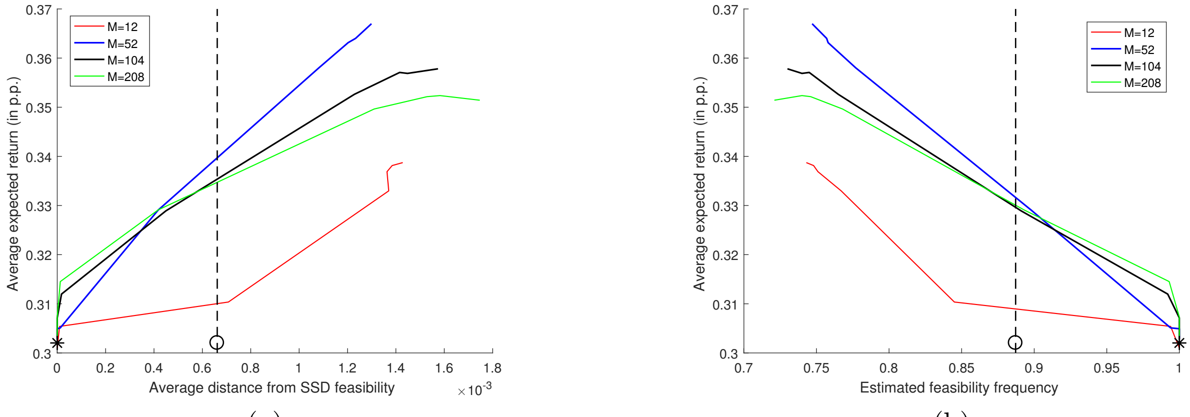


**Figure 6** Statistics of the feasibility on validation data. (a) presents the mean (solid), 10-th, and 90-th percentiles (bars) of the out-of-sample distance from SSD feasibility. (b) presents the estimated feasibility frequency (with 90% confidence intervals) on validation data as a function of $\epsilon$ for $M \in \{12, 52, 104, 208\}$ empirical weeks. The dashed lines identify an acceptable level of performance that is based on a re-sampled version of the reference return's distribution.



(a) (b)

**Figure 7** Bi-objective performance curves for strategies produced by the DRSSDCP on validation data. (a) presents the average expected return vs. the average distance from SSD feasibility. (b) presents the average expected return vs. the estimated feasibility frequency. In both figures, the star and the circle indicate respectively the performance achieved by the reference portfolio and a re-sampled version of the reference return's distribution.

**8.2.4. Out-of-Sample Performance of DRSSDCP Solutions** In this section, we evaluate the out-of-sample performance (i.e. using the out-of-sample data) of the DRSSDCP model that was calibrated in Section 8.2.3, i.e. with $M = 52$ and $\epsilon = 0.01$. To be consistent with the cross-validation procedure, these out-of-sample experiments still involve $m = 5$ and will consider 12 000 runs that each involve blocks of train ($M = 52$ weeks) and test ($M' = 26$ weeks) data obtained again by picking a random subset of $m = 5$ companies and a random week in the 6 years out-of-sample period. The performance of DRSSDCP is compared to the performance of SAA, which uses the $M = 52$ most recent weekly return observations, and of the reference portfolio.

Table 3 compares the statistics of the out-of-sample performance for the three types of portfolios, including the expected return, standard deviation, 90% of CVaR, SSD distance and SSD feasibility. From Table 3, we observe that our DRSSDCP policy achieves the highest average expected return among all three type of portfolios. In terms of out-of-sample SSD feasibility, we see that the DRSSDCP portfolios has a significantly smaller average SSD distance than SAA and actually falls below the estimated acceptable threshold. Similar observations can be made in terms of feasibility frequency, namely the DRSSDCP portfolios were feasible with respect to the SSD constraint in 95.7% of the 12 000 runs. This out-of-sample feasibility performance again slightly exceeds the estimated acceptable threshold (94.2%). These results confirm that our calibrated DRSSDCP model performs very well out-of-sample when compared with an SAA approach and the reference portfolio.

In terms of statistical significance of our findings, we conducted one-side t-tests to verify if we could safely reject the hypothesis that each mean of the three statistics of interest (expected return,

**Table 3** The average performance for the SAA, DRSSDCP ($\epsilon = 0.01$) and reference policies in terms of expected return, standard deviation, 90% of CVaR, and SSD distance and feasibility frequency, in which the portfolio is rebalanced every 26 weeks.

| Descriptive statistics | SAA | DRSSDCP | Reference | Acceptable threshold |
|---|---|---|---|---|
| Average expected return (in p.p.) | 0.183 | 0.190 | 0.184 | - |
| Average standard deviation | 0.032 | 0.029 | 0.022 | - |
| Average CVaR(0.90) | 0.054 | 0.048 | 0.037 | - |
| Average SSD distance ($\times 10^{-3}$) | 0.486 | 0.088 | 0 | 0.225 |
| SSD feasibility frequency | 0.868 | 0.957 | 1 | 0.942 |

distance to feasibility, and actual feasibility) is worst for the DRSSDCP portfolios than for the portfolios obtained from SAA or the reference portfolio. For instance, the hypotheses comparing DRSSDCP to SAA with respect to the mean of expected returns can be described as follows:

$\mathcal{H}_0$ : the mean of expected returns for DRSSDCP portfolios is smaller than for SAA portfolios.

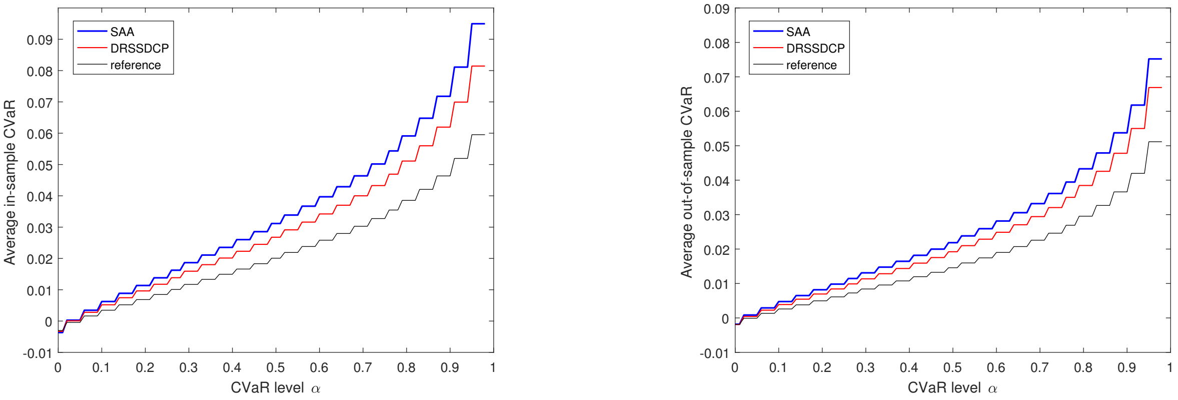
$\mathcal{H}_1$ : the mean of expected returns for DRSSDCP portfolios is larger than for SAA portfolios.

Table 4 reports the $p$-values for all six hypothesis tests. If we choose a significance level of 5%, then we see that all three hypothesis that compares to SAA can safely be rejected. We can also see that we can safely reject the hypothesis that the mean of expected returns for DRSSDCP portfolios is smaller that for the reference portfolio. On the other hand, it comes with no surprise that we can only accept the fact that the mean SSD distance and feasibility frequency are respectively larger and smaller for DRSSDCP than for the reference portfolio.

**Table 4** The $p$-values of the one-side t-tests describing the chance of falsely rejecting the hypothesis that DRSSDCP has a worse performance

| Different policies | SAA | Reference |
|---|---|---|
| Mean expected return | $10^{-4}$ | 0.05 |
| Mean SSD distance | 0 | 1 |
| SSD feasibility probability | 0 | 1 |

We conclude this section with one last set of results comparing the average CVaR profiles of the three approaches computed on the in-sample (on validation data) and out-of-sample (on test data) data sets. This is presented in Figure 8. One can observe that the average CVaR profile for the DRSSDCP portfolios dominates the profile of the SAA portfolios. When comparing the DRSSDCP profile to the profile of the reference portfolio, we notice that the latter performs better for strictly positive risk aversion levels ($\alpha$). This appears to be the price that DRSSDCP pays in order to improve by 0.006 percentage point the average expected return (see Table 3).



(a) (b)

**Figure 8** The average in-sample CVaR profile of the calibrated models (in (a)) and out-of-sample CVaR profile with 26 weekly returns (in (b)) as a function of the risk level $\alpha$ for the SAA, DRSSDCP ($\epsilon = 0.01$) and reference policies with the test data.

## 9. Concluding Remarks

In this paper, we present the first comprehensive study of a data-driven formulation of the DRSSDCP that hinges on using the Wasserstein ambiguity set proposed in Mohajerin Esfahani and Kuhn (2018). Our study includes summarizing two valuable statistical properties (namely finite sample guarantees and asymptotic consistency) of its solution in a data-driven context, identifying two tractable approximations and an exact solution algorithm, and performing an extensive set of numerical experiments with a special focus on the out-of-sample feasibility of the SSD constraint.

Methodologically speaking, we believe that our proposed results are flexible enough to straightforwardly be adapted to other versions of the DRSSDCP. While we already summarized how this can be done for the decomposable-DRSSDCP formulation, which is convenient when the information about the reference variable corresponds to a known distribution function, this could also straightforwardly be done for data-driven DRSSDCP that exploit a type-infinity (instead of type-1) Wasserstein ambiguity set (e.g., Gao and Kleywegt 2017, Bertsimas et al. 2021, Xie 2020). Future research should investigate how to explore the potential strategies that could be used to speed up our solution algorithm for large-scale problems (e.g., up to 1000 stocks and 500 empirical samples) or to address multivariate SSD constrained problems (e.g., Dentcheva and Ruszczyński 2009, Homem-de Mello and Mehrotra 2009, Hu et al. 2012).

On the empirical side, this paper demonstrated how, in a data-driven portfolio optimization problem, out-of-sample SSD feasibility can be improved by carefully tuning the level of robustification of the DRSSDCP without sacrificing much (if at all) in terms of expected return. We are now especially curious to see what might be the benefits of using the data-driven DRSSDCP/D-DRSSDCP in other fields of applications where it is natural to compare the performance of the decisions to a reference performance, e.g. EMS location problems such as in Noyan (2010) and Peng et al. (2020).

## Acknowledgments

The authors would like to thank the associate editor and the two anonymous reviewers for their useful and insightful comments that have significantly improved the paper. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada [Grant RGPIN-2016-05208], the Canada Research Chair program [Grant 950-230057], and Groupe d'études et de recherche en analyse des décisions. Finally, this research was enabled in part by support provided by Compute Canada (https://www.computecanada.ca/).

## References

AlAshery, M. K., Xiao, D., Qiao, W., 2019. Second-order stochastic dominance constraints for risk management of a wind power producer's optimal bidding strategy. IEEE Transactions on Sustainable Energy 11 (3), 1404–1413.

Armbruster, B., Delage, E., 2015. Decision making under uncertainty when preference information is incomplete. Management Science 61 (1), 111–128.

Armbruster, B., Luedtke, J., 2015. Models and formulations for multivariate dominance-constrained stochastic programs. IIE Transactions 47 (1), 1–14.

Bawa, V. S., 1982. Research bibliographystochastic dominance: A research bibliography. Management Science 28 (6), 698–712.

Ben-Tal, A., den Hertog, D., Vial, J.-P., 2015. Deriving robust counterparts of nonlinear uncertain inequalities. Mathematical Programming 149 (1-2), 265–299.

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

Bertsimas, D., Dunning, I., 2016. Multistage robust mixed-integer optimization with adaptive partitions. Operations Research 64 (4), 980–998.

Bertsimas, D., Shtern, S., Sturt, B., 2021. A data-driven approach for multi-stage linear optimization. forthcoming in Management Science, http://www.optimization-online.org/DB_FILE/2018/11/6907.pdf.

Bertsimas, D., Sim, M., 2004. The price of robustness. Operations Research 52 (1), 35–53.

Carrión, M., Gotzes, U., Schultz, R., 2009. Risk aversion for an electricity retailer with second-order stochastic dominance constraints. Computational Management Science 6 (2), 233–250.

Chen, Z., Jiang, J., 2018. Stability analysis of optimization problems with k th order stochastic and distributionally robust dominance constraints induced by full random recourse. SIAM Journal on Optimization 28 (2), 1396–1419.

Chen, Z., Kuhn, D., Wiesemann, W., 2018. Data-driven chance constrained programs over wasserstein balls. working draft, http://www.optimization-online.org/DB_FILE/2018/06/6671.pdf.

Delage, E., Guo, S., Xu, H., 2021. Shortfall risk models when information of loss function is incomplete. forthcoming in Operations Research, http://www.optimization-online.org/DB_FILE/2018/04/6593.pdf.

Delage, E., Kuhn, D., Wiesemann, W., 2019. Dice-sionmaking under uncertainty: When can a random decision reduce risk? Management Science 65 (7), 3282–3301.

Delage, E., Li, J. Y.-M., 2018. Minimizing risk exposure when the choice of a risk measure is ambiguous. Management Science 64 (1), 327–344.

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

Dentcheva, D., Martinez, G., 2012. Two-stage stochastic optimization problems with stochastic ordering constraints on the recourse. European Journal of Operational Research 219 (1), 1–8.

Dentcheva, D., Martinez, G., Wolfhagen, E., 2016. Augmented lagrangian methods for solving optimization problems with stochastic-order constraints. Operations Research 64 (6), 1451–1465.

Dentcheva, D., Ruszczynski, A., 2003. Optimization with stochastic dominance constraints. SIAM Journal on Optimization 14 (2), 548–566.

Dentcheva, D., Ruszczyński, A., 2009. Optimization with multivariate stochastic dominance constraints. Mathematical Programming 117 (1-2), 111–127.

Dentcheva, D., Ruszczyński, A., 2010. Robust stochastic dominance and its application to risk-averse optimization. Mathematical Programming 123 (1), 85–100.

Dentcheva, D., Wolfhagen, E., 2015. Optimization with multivariate stochastic dominance constraints. SIAM Journal on Optimization 25 (1), 564–588.

Dentcheva, D., Wolfhagen, E., 2016. Two-stage optimization problems with multivariate stochastic order constraints. Mathematics of Operations Research 41 (1), 1–22.

Gao, R., Kleywegt, A. J., 2016. Distributionally robust stochastic optimization with wasserstein distance. arXiv preprint arXiv:1604.02199.

Gao, R., Kleywegt, A. J., 2017. Distributionally robust stochastic optimization with dependence structure. arXiv preprint arXiv:1701.04200.

Guo, S., Xu, H., 2019. Distributionally robust shortfall risk optimization model and its approximation. Mathematical Programming 174 (1), 473–498.

Guo, S., Xu, H., Zhang, L., 2017. Probability approximation schemes for stochastic programs with distributionally robust second-order dominance constraints. Optimization Methods and Software 32 (4), 770–789.

Hadar, J., Russell, W. R., 1969. Rules for ordering uncertain prospects. The American Economic Review 59 (1), 25–34.

Hadjiyiannis, M. J., Goulart, P. J., Kuhn, D., 2011. A scenario approach for estimating the suboptimality of linear decision rules in two-stage robust optimization. In: Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on. IEEE, pp. 7386–7391.

Haskell, W. B., Fu, L., Dessouky, M., 2016. Ambiguity in risk preferences in robust stochastic optimization. European Journal of Operational Research 254 (1), 214–225.

Haskell, W. B., Shanthikumar, J. G., Shen, Z. M., 2017. Primal-dual algorithms for optimization with stochastic dominance. SIAM Journal on Optimization 27 (1), 34–66.

Homem-de Mello, T., Mehrotra, S., 2009. A cutting-surface method for uncertain linear programs with polyhedral stochastic dominance constraints. SIAM Journal on Optimization 20 (3), 1250–1273.

Hu, J., Homem-de Mello, T., Mehrotra, S., 2011. Risk-adjusted budget allocation models with application in homeland security. IIE Transactions 43 (12), 819–839.

Hu, J., Homem-de Mello, T., Mehrotra, S., 2012. Sample average approximation of stochastic dominance constrained programs. Mathematical Programming 133 (1-2), 171–201.

Huang, R. J., Tzeng, L. Y., Zhao, L., 2020. Fractional degree stochastic dominance. Management Science.

Ji, R., Lejeune, M. A., 2021. Data-driven optimization of reward-risk ratio measures. INFORMS Journal on Computing 33 (3), 1120–1137.

Kleywegt, A. J., Shapiro, A., Homem-de Mello, T., 2002. The sample average approximation method for stochastic discrete optimization. SIAM Journal on Optimization 12 (2), 479–502.

Kozmík, K., 2019. Robust approaches in portfolio optimization with stochastic dominance. Master Thesis in Univerzita Karlova.

Liesiö, J., Xu, P., Kuosmanen, T., 2020. Portfolio diversification based on stochastic dominance under incomplete probability information. European Journal of Operational Research.

Lizyayev, A., Ruszczyński, A., 2012. Tractable almost stochastic dominance. European Journal of Operational Research 218 (2), 448–455.

Long, D. Z., Sim, M., Zhou, M., 2021. Robust satisficing. working draft, `http://www.optimization-online.org/DB_FILE/2019/11/7456.pdf`.

Luedtke, J., 2008. New formulations for optimization under stochastic dominance constraints. SIAM Journal on Optimization 19 (3), 1433–1450.

Mohajerin Esfahani, P., Kuhn, D., 2018. Data-driven distributionally robust optimization using wasserstein metric: Performance guarantees and tractable reformulations. Mathematical Programming 171 (1-2), 115–166.

Montes, I., Miranda, E., Montes, S., 2014. Stochastic dominance with imprecise information. Computational Statistics & Data Analysis 71, 868–886.

Müller, A., Scarsini, M., Tsetlin, I., Winkler, R. L., 2017. Between first-and second-order stochastic dominance. Management Science 63 (9), 2933–2947.

Müller, A., Stoyan, D., 2002. Comparison methods for stochastic models and risks. Vol. 389. Wiley New York.

Noyan, N., 2010. Alternate risk measures for emergency medical service system design. Annals of Operations Research 181 (1), 559–589.

Noyan, N., 2018. Risk-averse stochastic modeling and optimization. In: INFORMS TutORials in Operations Research: Recent Advances in Optimization and Modeling of Contemporary Problems. INFORMS, pp. 221–254.

Noyan, N., Rudolf, G., 2013. Optimization with multivariate conditional value-at-risk constraints. Operations Research 61 (4), 990–1013.

Noyan, N., Rudolf, G., 2018. Optimization with stochastic preferences based on a general class of scalarization functions. Operations Research 66 (2), 463–486.

Peng, C., Delage, E., Li, J., 2020. Probabilistic envelope constrained multiperiod stochastic emergency medical services location model and decomposition scheme. Transportation Science 54 (6), 1471–1494.

Post, T., Kopa, M., 2017. Portfolio choice based on third-degree stochastic dominance. Management Science 63 (10), 3381–3392.

Postek, K., den Hertog, D., 2016. Multistage adjustable robust mixed-integer optimization via iterative splitting of the uncertainty set. INFORMS Journal on Computing 28 (3), 553–574.

Quiggin, J., 1993. Generalized Expected Utility Theory: The Rank Dependent Model. Springer Science & Business Media.

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

Roman, D., Mitra, G., Zverovich, V., 2013. Enhanced indexation based on second-order stochastic dominance. European Journal of Operational Research 228 (1), 273–281.

Rudolf, G., Ruszczyński, A., 2008. Optimization problems with second order stochastic dominance constraints: duality, compact formulations, and cut generation methods. SIAM Journal on Optimization 19 (3), 1326–1343.

Ruszczyński, A., 2013. Advances in risk-averse optimization. In: INFORMS TutORials in Operations Research: Theory Driven by Influential Applications. INFORMS, pp. 168–190.

Sehgal, R., Mehra, A., 2020. Robust portfolio optimization with second order stochastic dominance constraints. Computers & Industrial Engineering, 106396.

Sion, M., et al., 1958. On general minimax theorems. Pacific Journal of mathematics 8 (1), 171–176.

Wiesemann, W., Kuhn, D., Sim, M., 2014. Distributionally robust convex optimization. Operations Research 62 (6), 1358–1376.

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

Xie, W., 2021. On distributionally robust chance constrained programs with wasserstein distance. Mathematical Programming 186 (1), 115–155.

Zarif, M., Javidi, M. H., Ghazizadeh, M. S., 2012. Self-scheduling of large consumers with second-order stochastic dominance constraints. IEEE Transactions on Power Systems 28 (1), 289–299.

Zhang, L., Homem-de Mello, T., 2016. An optimal path model for the risk-averse traveler. Transportation Science 51 (2), 518–535.

Zhang, L., Homem-de Mello, T., 2017. An optimal path model for the risk-averse traveler. Transportation Science 51 (2), 518–535.

Zhao, C., Guan, Y., 2018. Data-driven risk-averse stochastic optimization with wasserstein metric. Operations Research Letters 46 (2), 262–267.

## Appendix A: Proofs

### A.1. Proof of Theorem 1

Our proof is straightforward since the properties of maximal ambiguity indecisiveness and ambiguity monotonicity are respectively responsible for the "if" and "only if" part of the statement: $X \succeq Y \Leftrightarrow \forall \mathbb{P} \in \mathcal{P},\, F_X^{\mathbb{P}} \succeq F_Y^{\mathbb{P}}$.

The second part of our claim is simple to derive based on the representation. Firstly, if $X \succeq Y \succeq Z$, then for all $\mathbb{P} \in \mathcal{P}$ we have that:

$$F_X^{\mathbb{P}} \succeq F_Y^{\mathbb{P}} \succeq F_Z^{\mathbb{P}}.$$

If $\succeq$ is transitive on $\mathcal{U}$ and distribution based, then this implies that $F_X^{\mathbb{P}} \succeq F_Z^{\mathbb{P}}$ for all $\mathbb{P} \in \mathcal{P}$, which lets us conclude that $X \succeq Z$. Secondly, we have that for all $\mathbb{P} \in \mathcal{P}$, $F_X^{\mathbb{P}} = F_X^{\mathbb{P}}$ so that if $\succeq$ is reflexive on $\mathcal{U}$ and distribution based, then $F_X^{\mathbb{P}} \succeq F_X^{\mathbb{P}}$ for all $\mathbb{P} \in \mathcal{P}$, hence $X \succeq X$. □

### A.2. Proof of Proposition 1

Based on its definition, the Wasserstein ambiguity set $\mathcal{P}_{\mathrm{W}}^{r}(\hat{\mathbb{P}}, \epsilon)$ reduces to the singleton $\{\hat{\mathbb{P}}\}$ when $\epsilon = 0$. Therefore, it is easy to verify that DRSSDCP (4) can reduce to the SDCP2 with $h(\boldsymbol{x}) := \boldsymbol{c}^\top \boldsymbol{x}$ and $\mathbb{P} := \hat{\mathbb{P}}$.

Next, we obtain a finite linear programming reformulation by exploiting the fact that $\hat{\mathbb{P}}$ is discrete and finite. In this context, one can employ Proposition 3.2 in Dentcheva and Ruszczynski (2003) to equivalently replace $t \in \mathbb{R}$ with $t \in \{t_1, \ldots, t_M\}$, where each $t_i := f_0(\hat{\boldsymbol{\xi}}_i)$. By introducing epigraph variables, we further obtain that DRSSDCP, i.e., SDCP2, reduces to the following problem:

$$\begin{aligned}
&\underset{\boldsymbol{x}\in\mathcal{X},\boldsymbol{s}\geq 0}{\text{minimize}}\ \boldsymbol{c}^\top\boldsymbol{x} \\
&\text{subject to}\ \frac{1}{M}\sum_{k\in[M]} s_{ik} \leq \frac{1}{M}\sum_{k\in[M]}(t_i - f_0(\hat{\boldsymbol{\xi}}_k))^+ && \forall i\in[M] \\
&\qquad f(\boldsymbol{x},\hat{\boldsymbol{\xi}}_k) + s_{ik} \geq t_i && \forall i\in[M], k\in[M].
\end{aligned}$$

Finally, under Assumption 2 and $\mathcal{X}$ polyhedral, this optimization model further reduces to the following linear program:

$$\begin{aligned}
&\underset{\boldsymbol{x}\in\mathcal{X},\boldsymbol{s}\geq 0}{\text{minimize}}\ \boldsymbol{c}^\top\boldsymbol{x} \\
&\text{subject to}\ \frac{1}{M}\sum_{k\in[M]} s_{ik} \leq \frac{1}{M}\sum_{k\in[M]}(t_i - f_0(\hat{\boldsymbol{\xi}}_k))^+ && \forall i\in[M] \\
&\qquad \boldsymbol{a}_n(\boldsymbol{x})^\top\hat{\boldsymbol{\xi}}_k + b_n(\boldsymbol{x}) + s_{ik} \geq t_i && \forall n\in[N], \forall i\in[M], k\in[M]. \quad \square
\end{aligned}$$

### A.3. Proof of Proposition 2

Our first step consists in proving the reduction of DRSSDCP to the DFSDCP (5) under $\mathcal{P}^r_{\mathrm{W}}(\hat{\mathbb{P}},\infty)$. According to the definition of $\mathcal{P}^r_{\mathrm{W}}(\hat{\mathbb{P}},\epsilon)$, if $\epsilon=\infty$, then $d_{\mathrm{W}}(\mathbb{P},\hat{\mathbb{P}})\leq\epsilon$ always holds. Thus, the Wasserstein ambiguity set $\mathcal{P}^r_{\mathrm{W}}(\hat{\mathbb{P}},\epsilon)$ reduces $\mathcal{M}(\Xi) := \{\mathbb{P} \mid \mathbb{P}(\boldsymbol{\xi}\in\Xi)=1\}$.

We firstly prove that constraint (4b) implies constraint (5b). If $f(\boldsymbol{x},\boldsymbol{\xi}) \succeq^{\mathbb{P}}_{(2)} f_0(\boldsymbol{\xi}),\ \forall\mathbb{P}\in\mathcal{M}(\Xi)$ is satisfied, then for any $\bar{\boldsymbol{\xi}}\in\Xi$, then $f(\boldsymbol{x},\boldsymbol{\xi}) \succeq^{\delta_{\bar{\xi}}}_{(2)} f_0(\boldsymbol{\xi})$, where $\delta_{\bar{\xi}}$ is the Dirac function that puts all the weight on $\bar{\boldsymbol{\xi}}$. Based on the representation of SSD in Lemma 1, for each $\bar{\boldsymbol{\xi}}\in\Xi$, we have:

$$\mathbb{E}_{\delta_{\bar{\xi}}}[(t-f(\boldsymbol{x},\bar{\boldsymbol{\xi}}))^+] \leq \mathbb{E}_{\delta_{\bar{\xi}}}[(t-f_0(\bar{\boldsymbol{\xi}}))^+],\ \forall t\in\mathbb{R} \quad\Leftrightarrow\quad (t-f(\boldsymbol{x},\bar{\boldsymbol{\xi}}))^+ \leq (t-f_0(\bar{\boldsymbol{\xi}}))^+,\ \forall t\in\mathbb{R}.$$

We can now show by contradiction that this constraint implies constraint (5b). Specifically, let there exist a $\bar{\boldsymbol{\xi}}\in\Xi$ such that $f(\boldsymbol{x},\bar{\boldsymbol{\xi}}) < f_0(\bar{\boldsymbol{\xi}})$ hold, then there must be a $\bar{t}$ for which $f(\boldsymbol{x},\bar{\boldsymbol{\xi}}) < \bar{t} < f_0(\bar{\boldsymbol{\xi}})$, which also implies that $\bar{t} - f_0(\bar{\boldsymbol{\xi}}) < 0 < \bar{t} - f(\boldsymbol{x},\bar{\boldsymbol{\xi}})$. Thus we have that:

$$\left(\bar{t} - f_0(\bar{\boldsymbol{\xi}})\right)^+ = 0 < \bar{t} - f(\boldsymbol{x},\bar{\boldsymbol{\xi}}) = \left(\bar{t} - f(\boldsymbol{x},\bar{\boldsymbol{\xi}})\right)^+ .$$

This contradicts the fact that $\left(t - f(\boldsymbol{x},\bar{\boldsymbol{\xi}})\right)^+ \leq \left(t - f_0(\bar{\boldsymbol{\xi}})\right)^+$ for all $t\in\mathbb{R}$ and all $\boldsymbol{\xi}\in\Xi$. Thus, we must have that $f(\boldsymbol{x},\bar{\boldsymbol{\xi}}) \geq f_0(\bar{\boldsymbol{\xi}}),\ \forall\bar{\boldsymbol{\xi}}\in\Xi$.

On the other hand, if constraint (5b) holds, then for all $\mathbb{P}\in\mathcal{M}(\Xi)$, we have $\mathbb{P}(f(\boldsymbol{x},\boldsymbol{\xi})\leq t) \leq \mathbb{P}(f_0(\boldsymbol{\xi})\leq t),\ \forall t\in\mathbb{R}$. In other words, $f(\boldsymbol{x},\boldsymbol{\xi})$ stochastically dominates $f_0(\boldsymbol{\xi})$ in the first-order for all $\mathbb{P}\in\mathcal{M}(\Xi)$, which is known to imply that the same dominance hold in the second-order. We can conclude that (5b) implies (4b), so that the two constraints are equivalent and DRSSDCP with $\mathcal{P}^r_{\mathrm{W}}(\hat{\mathbb{P}},\infty)$ reduces to the DFSDCP (5).

Next, we derive a linear programming formulation under mild conditions. Under assumptions 1 and 2, constraint (5b) can be further rewritten as

$$\min_{n\in[N]} \boldsymbol{a}_n(\boldsymbol{x})^\top \boldsymbol{\xi} + b_n(\boldsymbol{x}) \geq \min_{n'\in[N]} {\boldsymbol{a}_n^0}^\top \boldsymbol{\xi} + b_n^0 \quad \forall \boldsymbol{\xi} \in \Xi,$$

which can be reformulated as

$$\forall\, n \in [N],\, \min_{\boldsymbol{\xi}\in\Xi} \max_{\boldsymbol{\theta}\in\mathbb{R}_+^N : \sum_{j=1}^N \theta_j^n = 1} \boldsymbol{a}_n(\boldsymbol{x})^\top \boldsymbol{\xi} + b_n(\boldsymbol{x}) - \sum_{j\in[N]} \theta_n^j ({\boldsymbol{a}_n^0}^\top \boldsymbol{\xi} + b_n^0) \geq 0.$$

Applying Sion's minimax theorem (see Sion et al. (1958)), we obtain an equivalent condition:

$$\forall\, n \in [N],\, \max_{\boldsymbol{\theta}\in\mathbb{R}_+^N : \sum_{j\in[N]} \theta_j = 1} \min_{\boldsymbol{\xi}\in\Xi} \boldsymbol{a}_n(\boldsymbol{x})^\top \boldsymbol{\xi} + b_n(\boldsymbol{x}) - \sum_{j\in[N]} \theta_j ({\boldsymbol{a}_n^0}^\top \boldsymbol{\xi} + b_n^0) \geq 0.$$

Based on the Fenchel duality (see Ben-Tal et al. (2015)), we can show that the condition can be rewritten as:

$$\exists\, \boldsymbol{\theta} \in \mathbb{R}_+^{N\times N},\, \forall\, n \in [N],\, \sum_{j\in[N]} \theta_n^j = 1 \quad \& \quad -\delta\left(\sum_{j\in[N]} \theta_n^j \boldsymbol{a}_n^0 - \boldsymbol{a}_n(\boldsymbol{x}) \,\middle|\, \Xi\right) - \sum_{j\in[N]} \theta_n^j b_n^0 + b_n(\boldsymbol{x}) \geq 0.$$

Thus, problem (5) is equivalent to the following convex optimization problem:

$$\begin{aligned} &\underset{\boldsymbol{x}\in\mathcal{X},\boldsymbol{\theta}\geq 0}{\text{minimize}}\ \boldsymbol{c}^\top \boldsymbol{x} \\ &\text{subject to } -\delta\left(\sum_{j\in[N]} \theta_n^j \boldsymbol{a}_n^0 - \boldsymbol{a}_n(\boldsymbol{x}) \,\middle|\, \Xi\right) - \sum_{j\in[N]} \theta_n^j b_n^0 + b_n(\boldsymbol{x}) \geq 0 \quad \forall n \in [N] \\ &\qquad\qquad \sum_{j\in[N]} \theta_n^j = 1 \qquad \forall n \in [N], \end{aligned}$$

where $\delta(\cdot)$ represents the support function and $\Xi$ is the support set. Moreover, it is easy to verify that it can be reformulated as a linear program if $\Xi$ has a linear programming representation support function (i.e. is polyhedral), and if $\mathcal{X}$ is polyhedral. □

### A.4. Proof of Proposition 3

Our proof mainly employs the result of finite sample guarantee in Theorem 3.5 of Mohajerin Esfahani and Kuhn (2018). Based on Theorem 3.5 in Mohajerin Esfahani and Kuhn (2018), and suppose that Assumption 1 holds and that each observations in $\{\hat{\boldsymbol{\xi}}_i\}_{i=1}^M$ are drawn i.i.d. from some $\bar{\mathbb{P}}$, for a given $\beta \in (0,1)$, we obtain that $\mathcal{P}_{\mathrm{W}}^1(\hat{\mathbb{P}}, \epsilon_M(\beta)$ is known to contain the true distribution $\bar{\mathbb{P}}$ with high probability $1-\beta$, where

$$\epsilon_M(\beta) := \begin{cases} \left(\dfrac{\log(c_1\beta^{-1})}{c_2 M}\right)^{1/\max(m,2)} & \text{if } M \geq \dfrac{\log(c_1\beta^{-1})}{c_2} \\ \left(\dfrac{\log(c_1\beta^{-1})}{c_2 M}\right)^{1/a} & \text{otherwise}\,, \end{cases}$$

and where $c_1$, $c_2$, and $a > 1$ are known positive constants.

Let $\hat{\boldsymbol{x}}_M$ be the optimal solution of the DRSSDCP with ambiguity set $\mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}}, \epsilon_M(\beta))$. If $\bar{\mathbb{P}}$ is in $\mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}}, \epsilon_M(\beta))$, then necessarily $\hat{\boldsymbol{x}}_M$ satisfies the SSD constraint over $\bar{\mathbb{P}}$. The probability that $\hat{\boldsymbol{x}}_M$ satisfies the SSD constraint over $\bar{\mathbb{P}}$ is therefore larger than $1-\beta$, given the above statistical property of $\mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}}, \epsilon_M(\beta))$. □

## A.5. Proof of Proposition 4

We start our proof by rewriting the $\phi$-DRSSDCP based on Section 5.3, as follows:

$$\begin{aligned}
&[\phi\text{-DRSSDCP}] \quad \underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}} \; \boldsymbol{c}^\top \boldsymbol{x} && (18a)\\
&\qquad\qquad \text{subject to } \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x},\boldsymbol{\xi},t)\right] \le \phi \qquad \forall t \in \bar{\mathcal{T}}, \forall \mathbb{P} \in \mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}}, \epsilon), && (18b)
\end{aligned}$$

where $g(\boldsymbol{x},\boldsymbol{\xi},t) := (t - f(\boldsymbol{x},\boldsymbol{\xi}))^+ - (t - f_0(\boldsymbol{\xi}))^+$ and $\bar{\mathcal{T}} := [t_{min}, t_{max}]$, with $t_{min} := \inf_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{\xi})$, and $t_{max} := \sup_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{\xi})$.

Similarly, we can also rewrite the $\phi$-SDCP2 in the form of

$$\begin{aligned}
&[\phi\text{-SDCP2}] \quad \underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}} \; \boldsymbol{c}^\top \boldsymbol{x} && (19a)\\
&\qquad\qquad \text{subject to } \mathbb{E}_{\bar{\mathbb{P}}}\left[g(\boldsymbol{x},\boldsymbol{\xi},t)\right] \le \phi \qquad \forall t \in \bar{\mathcal{T}}. && (19b)
\end{aligned}$$

Let $f^M, \boldsymbol{x}_M, \mathcal{X}_M$ be the optimal value, optimal solution and optimal solution set of the $\phi$-DRSSDCP respectively. Let $f^*, \boldsymbol{x}^*, \mathcal{X}^*$ be the optimal value, optimal solution and optimal solution set of the $\phi$-SDCP2 respectively.

To clarify presentation, we define

$$\bar{v}(\boldsymbol{x}) := \sup_{t\in\bar{\mathcal{T}}} \mathbb{E}_{\bar{\mathbb{P}}}\left[g(\boldsymbol{x},\boldsymbol{\xi},t)\right] - \phi \;\text{ and }\; v^M(\boldsymbol{x}) := \sup_{\mathbb{P}\in\mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}},\epsilon_M(\beta_M))} \sup_{t\in\bar{\mathcal{T}}} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x},\boldsymbol{\xi},t)\right] - \phi,$$

respectively.

Now we can easily establish the following two lemmas.

Lemma 2. *There exists an optimal solution $\boldsymbol{x}'^* \in \mathcal{X}^*$ such that for any given $\tau > 0$, there is a $\boldsymbol{x} \in \mathcal{X}$, with $\|\boldsymbol{x} - \boldsymbol{x}'^*\|_2 \le \tau$, $\bar{v}(\boldsymbol{x}) < 0$.*

. **Proof of Lemma 2** This results is trivial if there exists a $\boldsymbol{x}'^* \in \mathcal{X}^*$ such that $\bar{v}(\boldsymbol{x}^*) < 0$. In the case where this does not apply, one can take any $\boldsymbol{x}'^* \in \mathcal{X}^*$ (with $\bar{v}(\boldsymbol{x}^*) = 0$), and use the feasible solution that satisfies Slater's condition, i.e. some $\bar{\boldsymbol{x}} \in \mathcal{X}$ that satisfies $\bar{v}(\bar{\boldsymbol{x}}) < 0$, to identify an $\boldsymbol{x}$ with the required property. Namely, for any $\tau$, one can identify a convex combination of $\boldsymbol{x}'^*$ and $\bar{\boldsymbol{x}}$, i.e. $\boldsymbol{x}_\theta := \theta\bar{\boldsymbol{x}} + (1-\theta)\boldsymbol{x}'^*$ with $0 < \theta \le 1$, such that $\|\boldsymbol{x}_\theta - \boldsymbol{x}'^*\|_2 \le \tau$. By convexity of $\mathcal{X}$ and $\bar{v}(\cdot)$, we necessarily have that:

$$\bar{v}(\boldsymbol{x}_\theta) = \bar{v}(\theta\bar{\boldsymbol{x}} + (1-\theta)\boldsymbol{x}'^*) \le \theta\bar{v}(\bar{\boldsymbol{x}}) + (1-\theta)\bar{v}(\boldsymbol{x}'^*) = \theta\bar{v}(\bar{\boldsymbol{x}}) < 0\,,$$

where we exploit the fact that $\bar{v}(\boldsymbol{x}'^*) = 0$, $\theta > 0$ and $\bar{v}(\bar{\boldsymbol{x}}) < 0$. □

LEMMA 3. *Given that Assumption 2 is satisfied, for any given $\boldsymbol{x} \in \mathcal{X}$, the function $g(\boldsymbol{x}, \boldsymbol{\xi}, t)$ is $L_\xi$-Lipschitz continuous in $\boldsymbol{\xi}$ with respect to the $\ell_p$-norm where:*

$$L_\xi := \sup_{\boldsymbol{x}\in\mathcal{X}} \max_{n\in[N]} \|\boldsymbol{a}_n(\boldsymbol{x})\|_{p*} + \max_{n\in[N]} \|\boldsymbol{a}_n^0\|_{p*},$$

*with $\|\cdot\|_{p*}$ as the dual of the $\ell_p$-norm. Moreover, it is $L_x$-Lipschitz continuous in $\boldsymbol{x}$ with respect to norm $\|\cdot\|_2$ for all $\boldsymbol{\xi} \in \Xi$, where:*

$$L_x := \sup_{\boldsymbol{\xi}\in\Xi} \max_{n\in[N]} \|\bar{\boldsymbol{a}}_n(\boldsymbol{\xi})\|_2,$$

*where we let $\bar{\boldsymbol{a}}_n(\boldsymbol{\xi})^T\boldsymbol{x} + \bar{b}_n(\boldsymbol{\xi})$ be the affine representation of $\boldsymbol{a}_n(\boldsymbol{x})^T\boldsymbol{\xi} + b_n(\boldsymbol{x})$ in $\boldsymbol{x}$.*

. **Proof of Lemma 3** First, we have that, based on the definition of $f(\boldsymbol{x}, \boldsymbol{\xi})$ and $f_0(\boldsymbol{\xi})$, both are Lipschitz continuous with constants $\max_{n\in[N]} \|\boldsymbol{a}_n(\boldsymbol{x})\|_{p*}$ and $\max_{n\in[N]} \|\boldsymbol{a}_n^0(\boldsymbol{x})\|_{p*}$ respectively. Focusing on $(t - f(\boldsymbol{x}, \boldsymbol{\xi}))^+$, it is clear that the Lipschitz constant remains below $\max_{n\in[N]} \|\boldsymbol{a}_n(\boldsymbol{x})\|_{p*}$ since the function takes the maximum between 0 and the difference between $t$ and $f(\boldsymbol{x}, \boldsymbol{\xi})$. The same argument applies for $(t - f_0(\boldsymbol{\xi}))^+$ to verify that its Lipschitz constant is below $\max_{n\in[N]} \|\boldsymbol{a}_n^0(\boldsymbol{x})\|_{p*}$. We then use the fact that the Lipschitz constant of a sum of functions is smaller then the sum of their Lipschitz constants. A similar argument applies for obtaining $L_x$. □

We first prove that $\bar{\mathbb{P}}^\infty$-almost surely we have that $f^M \to f^*$ as $M \to \infty$. We will then derive the implications regarding the convergence of $\boldsymbol{x}_M$ to $\mathcal{X}^*$.

Using similar arguments as used in proving Proposition 3, we have that, for $\boldsymbol{x}_M \in \mathcal{X}_M$, $\boldsymbol{x}_M$ satisfies the relaxed SDC in $\phi$-SDCP2 with a probability of at least $1 - \beta_M$. Therefore, we have

$$\bar{\mathbb{P}}^M \left\{ f^* \leq f^M \right\} = \bar{\mathbb{P}}^M \left\{ f^* \leq \boldsymbol{c}^\top \boldsymbol{x}_M \right\} \geq \bar{\mathbb{P}}^M \left\{ \mathbb{E}_{\bar{\mathbb{P}}}\left[g(\boldsymbol{x}_M, \boldsymbol{\xi}, t)\right] \leq \phi,\, \forall t \in \bar{\mathcal{T}} \right\} \geq 1 - \beta_M.$$

Since $\sum_{M=1}^\infty \beta_M < \infty$ and based on the Borel-Cantelli Lemma, we further have that

$$\bar{\mathbb{P}}^\infty \left\{ f^* \leq f^M \text{ for all sufficiently large } M \right\} = 1.$$

To reach our objective, we are left with proving that

$$\bar{\mathbb{P}}^\infty \left\{ f^* \geq f^M \text{ for all sufficiently large } M \right\} = 1.$$

To do so, we first show that for all $\boldsymbol{x} \in \mathcal{X}$:

$$\limsup_{M\to\infty} v^M(\boldsymbol{x}) = \bar{v}(\boldsymbol{x})$$

$\bar{\mathbb{P}}^\infty$-almost surely holds.

Given any $\boldsymbol{x} \in \mathcal{X}$ and any sequence $\delta_M > 0$ such that $\lim_{M\to\infty} \delta_M = 0$, let $\hat{\mathbb{Q}}_M \in \mathcal{P}_{\mathrm{W}}^1(\hat{\mathbb{P}}_M, \epsilon_M(\beta_M))$ be a $\delta_M$-optimal worst-case distribution with

$$\sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{W}}^1(\hat{\mathbb{P}}_M, \epsilon_M(\beta_M))} \sup_{t\in\bar{\mathcal{T}}} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x}, \boldsymbol{\xi}, t)\right] \leq \sup_{t\in\bar{\mathcal{T}}} \mathbb{E}_{\hat{\mathbb{Q}}_M}\left[g(\boldsymbol{x}, \boldsymbol{\xi}, t)\right] + \delta_M.$$

Such a probability measure exists given that

$$\sup_{\mathbb{P}\in\mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}}_M,\epsilon_M(\beta_M))} \sup_{t\in\bar{\mathcal{T}}} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x},\boldsymbol{\xi},t)\right] \le \sup_{\boldsymbol{\xi}\in\Xi} \sup_{t\in\bar{\mathcal{T}}} g(\boldsymbol{x},\boldsymbol{\xi},t) < \infty\,,$$

due to $\bar{\mathcal{T}}$ and $\Xi$ being bounded (see Assumption 1) and $g(\boldsymbol{x},\boldsymbol{\xi},t)$ being continuous in $\boldsymbol{\xi}$ and $t$.

Hence, for any given $\boldsymbol{x}\in\mathcal{X}$ we have

$$\begin{aligned}
\limsup_{M\to\infty} v^M(\boldsymbol{x}) &= \limsup_{M\to\infty} \sup_{\mathbb{P}\in\mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}}_M,\epsilon_M(\beta_M))} \sup_{t\in\bar{\mathcal{T}}} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x},\boldsymbol{\xi},t)\right] - \phi && (20)\\
&\le \limsup_{M\to\infty} \sup_{t\in\bar{\mathcal{T}}} \mathbb{E}_{\hat{\mathbb{Q}}_M}\left[g(\boldsymbol{x},\boldsymbol{\xi},t)\right] + \delta_M - \phi && (21)\\
&\le \limsup_{M\to\infty} \sup_{t\in\bar{\mathcal{T}}} \mathbb{E}_{\bar{\mathbb{P}}}\left[g(\boldsymbol{x},\boldsymbol{\xi},t)\right] + L_\xi d_{\mathrm{W}}(\bar{\mathbb{P}},\hat{\mathbb{Q}}_M) + \delta_M - \phi && (22)\\
&= \sup_{t\in\bar{\mathcal{T}}} \mathbb{E}_{\bar{\mathbb{P}}}\left[g(\boldsymbol{x},\boldsymbol{\xi},t)\right] - \phi \quad \bar{\mathbb{P}}^\infty\text{-almost surely} && (23)\\
&= \bar{v}(\boldsymbol{x}), && (24)
\end{aligned}$$

where the equality (20) is based on the definition of $v^M(\boldsymbol{x})$ and the inequality (21) is based on the $\delta_M$-optimal worst-case distribution $\hat{\mathbb{Q}}_M$. The inequality (22) follows from Theorem 3.2 in Mohajerin Esfahani and Kuhn (2018) and the equality (23) holds due to Lemma 3.7 in Mohajerin Esfahani and Kuhn (2018), which shows the almost sure convergence of any sequence $\hat{\mathbb{Q}}_M \in \mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}}_M,\epsilon_M(\beta_M))$ to $\bar{\mathbb{P}}$ under the Wasserstein metric if Assumption 1 holds and $\beta_M\in(0,1)$ satisfies $\sum_{M=1}^{\infty}\beta_M<\infty$ and $\lim_{M\to\infty}\epsilon_M(\beta_M)=0$.

Based on Lemma 2, for all $\tau>0$ there is a $\boldsymbol{x}_\tau\in\mathcal{X}$, with $\|\boldsymbol{x}_\tau-\boldsymbol{x}'^*\|_2\le\tau$ and $\bar{v}(\boldsymbol{x}_\tau)<0$. By passing to a sub-sequence if necessary, we may assume without loss of generality that $\lim\limits_{\tau\to 0}\boldsymbol{x}_\tau=\boldsymbol{x}'^*$. Then we can obtain that

$$\bar{\mathbb{P}}^\infty\left\{v^M(\boldsymbol{x}_\tau)\le 0 \ \text{ for all sufficiently large } M\right\}=1,$$

hence that

$$\bar{\mathbb{P}}^\infty\left\{\boldsymbol{c}^\top\boldsymbol{x}_\tau\ge f^M \ \text{ for all sufficiently large } M\right\}=1.$$

Finally, we have that

$$\bar{\mathbb{P}}^\infty\left\{f^*=\boldsymbol{c}^\top\boldsymbol{x}'^*=\lim_{\tau\to 0}\boldsymbol{c}^\top\boldsymbol{x}_\tau\ge\lim_{M\to\infty}f^M\right\}=1.$$

Therefore, so far we have proved that $\bar{\mathbb{P}}^\infty\{f^M=f^* \ \text{ for all sufficiently large } M\}=1$.

We can now turn to showing that $\boldsymbol{x}_M$ converges almost surely to $\mathcal{X}^*$ as $M$ goes to infinity. Specifically, we will demonstrate (by contradiction) that

$$\bar{\mathbb{P}}^\infty\{\lim_{M\to\infty}D(\mathcal{X}_M,\mathcal{X}^*)=0\}=1$$

where $D(\cdot,\cdot)$ denotes the Hausdorff distance between two sets, i.e., $D(A,B):=\max(\sup_{\boldsymbol{x}\in A}\operatorname{dist}(\boldsymbol{x},B),\ \sup_{\boldsymbol{x}\in B}\operatorname{dist}(\boldsymbol{x},A))$ with $\operatorname{dist}(\boldsymbol{x},B)=\inf\limits_{\boldsymbol{y}\in B}\|\boldsymbol{x}-\boldsymbol{y}\|$. Let us assume that

$\bar{\mathbb{P}}^\infty\{\lim_{M\to\infty} D(\mathcal{X}_M,\mathcal{X}^*)>0\}>0$. Looking at any realisation where $\lim_{M\to\infty} D(\mathcal{X}_M,\mathcal{X}^*)>0$, we have that there must exist a sequence $\boldsymbol{x}'_M\in\mathcal{X}_M$, such that $\lim_{M\to\infty}\mathrm{dist}(\boldsymbol{x}'_M,\mathcal{X}^*)>0$. Since we know that the feasible set $\mathcal{X}$ is convex and compact, by passing to a sub-sequence if necessary, we may assume that $\boldsymbol{x}'_M\to\boldsymbol{x}'\in\mathcal{X}$. So we have $\boldsymbol{x}'\notin\mathcal{X}^*$. Recall that $\boldsymbol{x}'_M\in\mathcal{X}_M$, that, with probability one, $v^M(\boldsymbol{x})$ converges pointwise to $\bar{v}(\boldsymbol{x})$, and that $\bar{v}(\boldsymbol{x})$ is continuous. Hence,

$$\begin{aligned}
\bar{v}(\boldsymbol{x}') &= \lim_{M\to\infty} v^M(\boldsymbol{x}')\\
&= \lim_{M\to\infty} v^M(\boldsymbol{x}'_M)+v^M(\boldsymbol{x}')-v^M(\boldsymbol{x}'_M)\\
&\le \lim_{M\to\infty} v^M(\boldsymbol{x}'_M)+\sup_{\mathbb{P}\in\mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}},\epsilon_M(\beta_M))}\ \sup_{t\in\bar{\mathcal{T}}}\mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x}',\boldsymbol{\xi},t)-g(\boldsymbol{x}'_M,\boldsymbol{\xi},t)\right]\\
&\le \lim_{M\to\infty} v^M(\boldsymbol{x}'_M)+\sup_{\mathbb{P}\in\mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}},\epsilon_M(\beta_M))}\ \sup_{t\in\bar{\mathcal{T}}}\mathbb{E}_{\mathbb{P}}\left[L_x\|\boldsymbol{x}'-\boldsymbol{x}'_M\|_2\right]\\
&= \lim_{M\to\infty} v^M(\boldsymbol{x}'_M)+L_x\|\boldsymbol{x}'-\boldsymbol{x}'_M\|_2\\
&= \lim_{M\to\infty} v^M(\boldsymbol{x}'_M)\le 0\,,
\end{aligned}$$

where the second inequality comes from Lemma 3.

We can therefore conclude that $\boldsymbol{x}'_M$ is a feasible solution of the $\phi$-SDCP2 and therefore that $\boldsymbol{c}^\top\boldsymbol{x}'>f^*$. These arguments point to:

$$0<\bar{\mathbb{P}}^\infty\{\lim_{M\to\infty} D(\mathcal{X}_M,\mathcal{X}^*)>0\}\le\bar{\mathbb{P}}^\infty\{\exists\boldsymbol{x}'\in\mathcal{X}/\mathcal{X}^*,\lim_{M\to\infty} f^M=\boldsymbol{c}^\top\boldsymbol{x}'>f^*\}=\bar{\mathbb{P}}^\infty\{\lim_{M\to\infty} f^M>f^*\}\,.$$

However, this contradicts the fact that $\mathbb{P}^\infty\{\lim_{M\to\infty} f^M=f^*\}=1$.

This completes our proof. □

### A.6. Proof of Proposition 5

We start our proof by presenting the following lemma, which exploits the compactness of $\Xi$ to present an equivalent representation of constraint (6b) where $t\in\mathbb{R}$ is restricted to a bounded interval $\bar{\mathcal{T}}$.

Lemma 4. *Under assumptions 1 and 2, then constraint* (6b) *is equivalent to*

$$\mathbb{E}_{\mathbb{P}}\left[(t-f(\boldsymbol{x},\boldsymbol{\xi}))^+\right]\le\mathbb{E}_{\mathbb{P}}\left[(t-f_0(\boldsymbol{\xi}))^+\right]\qquad\forall\mathbb{P}\in\mathcal{P}^1_W(\hat{\mathbb{P}},\epsilon),\ \forall t\in\bar{\mathcal{T}}\,,\tag{25}$$

*where* $\bar{\mathcal{T}}:=[t_{min},t_{max}]$, *with* $t_{min}:=\inf_{\boldsymbol{\xi}\in\Xi}f_0(\boldsymbol{\xi})$, *and* $t_{max}:=\sup_{\boldsymbol{\xi}\in\Xi}f_0(\boldsymbol{\xi})$.

. **Proof of Lemma 4** Our proof simply relies on verifying that this equivalence holds when $\mathbb{P}\in\mathcal{P}^1_{\mathrm{W}}$ is fixed. Since $\bar{\mathcal{T}}$'s definition is independent of $\mathbb{P}$, we will be able to conclude that the equivalence holds for all $\mathbb{P}\in\mathcal{P}^1_{\mathrm{W}}$.

When $\mathbb{P}\in\mathcal{P}^1_{\mathrm{W}}$ is fixed, Proposition 1.2 in Hu et al. (2012) states that

$$\mathbb{E}_{\mathbb{P}}\left[(t-f(\boldsymbol{x},\boldsymbol{\xi}))^+\right]\le\mathbb{E}_{\mathbb{P}}\left[(t-f_0(\boldsymbol{\xi}))^+\right],\ t\in\mathbb{R}$$

is equivalent to

$$\mathbb{E}_{\mathbb{P}}\left[(t-f(\boldsymbol{x},\boldsymbol{\xi}))^+\right] \leq \mathbb{E}_{\mathbb{P}}\left[(t-f_0(\boldsymbol{\xi}))^+\right],\, t \in \bar{\mathcal{T}}$$

as long as $f_0(\boldsymbol{\xi})$'s support is bounded by $\bar{\mathcal{T}}$. We are left with confirming that with probability one:

$$f_0(\boldsymbol{\xi}) \in [\inf_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{\xi}),\, \sup_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{\xi})] = [t_{min},\, t_{max}]\,,$$

where both $t_{min}$ and $t_{max}$ are finite since $f_0(\boldsymbol{\xi})$ is assumed piecewise linear (based on Assumption 2) and $\Xi$ is assumed compact (based on Assumption 1). □

Next, let us define $g(\boldsymbol{x},\boldsymbol{\xi},t) := (t-f(\boldsymbol{x},\boldsymbol{\xi}))^+ - (t-f_0(\boldsymbol{\xi}))^+$. We then give the following lemma that is later used for deriving the reformulation of constraint (25).

Lemma 5. *Suppose that Assumption 2 holds, for any $\boldsymbol{x} \in \mathbb{R}^m$, the function $g(\boldsymbol{x},\boldsymbol{\xi},t)$ is the maximum of jointly piecewise linear concave functions in $\boldsymbol{\xi}$ and $t$. Specifically, it can be represented as:*

$$g(\boldsymbol{x},\boldsymbol{\xi},t) := \max_{n\in[N+1]} g_n(\boldsymbol{x},\boldsymbol{\xi},t)\,,$$

*where*

$$g_n(\boldsymbol{x},\boldsymbol{\xi},t) := \min_{n'\in[N+1]} (\boldsymbol{a}^0_{n'} - \boldsymbol{a}_n(\boldsymbol{x}))^\top\boldsymbol{\xi} + b^0_{n'} - b_n(\boldsymbol{x}) - (c^0_{n'} - c_n)t$$

*where $a_{N+1} = b_{N+1} = c_{N+1} = a^0_{N+1} = b^0_{N+1} = c_{N+1} = 0$, and $c_n = c^0_n = 1$ for all $n \in [N]$.*

. **Proof of Lemma 5**: Based on Assumption 2, we can rewrite $g(\boldsymbol{x},\boldsymbol{\xi},t)$ as

$$\begin{aligned}
g(\boldsymbol{x},\boldsymbol{\xi},t) &= \left(t - \min_{n\in[N]} \boldsymbol{a}_n(\boldsymbol{x})^\top\boldsymbol{\xi} + b_n(\boldsymbol{x})\right)^+ - \left(t - \min_{n\in[N]} {\boldsymbol{a}^0_n}^\top\boldsymbol{\xi} + b^0_n\right)^+ \\
&= \max\left(0,\, \max_{n\in[N]} t - \boldsymbol{a}_n(\boldsymbol{x})^\top\boldsymbol{\xi} - b_n(\boldsymbol{x})\right) + \min\left(0,\, \min_{n\in[N]} -t + {\boldsymbol{a}^0_n}^\top\boldsymbol{\xi} + b^0_n\right) \\
&= \max_{n\in[N+1]} c_n t - \boldsymbol{a}_n(\boldsymbol{x})^\top\boldsymbol{\xi} - b_n(\boldsymbol{x}) + \min_{n'\in[N+1]} {\boldsymbol{a}^0_{n'}}^\top\boldsymbol{\xi} + b^0_{n'} - c^0_{n'}t \\
&= \max_{n\in[N+1]} g_n(\boldsymbol{x},\boldsymbol{\xi},t)\,. \quad \square
\end{aligned}$$

Finally, we move to derive the multistage robust optimization formulation of DRSSDCP under $\mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}},\epsilon)$ with $\epsilon \in (0,\infty)$. We know that $g(\boldsymbol{x},\boldsymbol{\xi},t)$ is a maximum of concave functions in $\boldsymbol{\xi}$ by Lemma 5 and that $\Xi$ is a non-empty convex and bounded set by Assumption 1. Under these conditions and by using the strong duality results in Mohajerin Esfahani and Kuhn (2018), for any given fixed $t \in \bar{\mathcal{T}}$, the worst-case expectation

$$\sup_{\mathbb{P}\in\mathcal{P}^1_{\mathrm{W}}(\hat{\mathbb{P}},\epsilon)} \mathbb{E}_{\mathbb{P}}[g(\boldsymbol{x},\boldsymbol{\xi},t)]$$

coincides with the optimal value of the following optimization problem:

$$\inf_{\lambda,\boldsymbol{q}} \;\lambda\epsilon + \frac{1}{M}\sum_{i\in[M]} q_i \tag{26a}$$

$$\text{subject to} \quad g(\boldsymbol{x},\boldsymbol{\xi},t) - \lambda\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i\| \le q_i \qquad \forall \boldsymbol{\xi}\in\Xi, i\in[M] \tag{26b}$$

$$\lambda \ge 0, \boldsymbol{q}\in\mathbb{R}^M. \tag{26c}$$

Therefore, let the infimum in (26) be denoted by $L(\boldsymbol{x},t)$, then DRSSDCP (4) is equivalent to the multistage robust optimization problem (8). Moreover, when Assumption 3 is satisfied, if the Wasserstein metric is a $\ell_1$-norm, constraint (26b) reduces to:

$$g_n(\boldsymbol{x},\boldsymbol{\xi},t) - \lambda\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i\|_1 \le q_i \quad \forall \boldsymbol{\xi}\in\Xi, n\in[N+1],\, i\in[M].$$

It further reduces to:

$$\min_{n'\in[N+1]} (\boldsymbol{a}^0_{n'} - \boldsymbol{a}_n(\boldsymbol{x}))^\top \boldsymbol{\xi} + b^0_{n'} - b_n(\boldsymbol{x}) - (c^0_{n'} - c_n)t - \lambda \sum_{j\in[m]} \boldsymbol{\nu}_j \le q_i \quad \forall \boldsymbol{\nu}\in\Gamma^i_\nu(\boldsymbol{\xi}),\, \boldsymbol{\xi}\in\Xi, n\in[N+1],\, i\in[M],$$

where $\Gamma^i_\nu(\boldsymbol{\xi}) := \{\boldsymbol{\nu}\in\mathbb{R}^m | -\boldsymbol{\nu} \le \boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i \le \boldsymbol{\nu}\}$. We finally, can introduce epigraph adversarial variables to obtain:

$$\eta - \boldsymbol{a}_n(\boldsymbol{x})^\top \boldsymbol{\xi} - b_n(\boldsymbol{x}) + c_n t - \lambda \sum_{j\in[m]} \boldsymbol{\nu}_j \le q_i \quad \forall \eta\in\Gamma_\eta(\boldsymbol{\xi},t),\, \boldsymbol{\nu}\in\Gamma_\nu(\boldsymbol{\xi}),\, \boldsymbol{\xi}\in\Xi, n\in[N+1],\, i\in[M]\,,$$

where $\Gamma_\eta(\boldsymbol{\xi},t) := \left\{\eta : \eta \le {\boldsymbol{a}^0_n}^\top \boldsymbol{\xi} + b^0_n - c^0_n t,\, \forall n\in[N+1]\right\}$. Thus obtaining the following second-stage robust linear program:

$$\begin{aligned} \inf_{\lambda,\boldsymbol{q}} \quad & \lambda\epsilon + \frac{1}{M}\sum_{i\in[M]} q_i \\ \text{subject to} \quad & \eta - \boldsymbol{a}_n(\boldsymbol{x})^\top \boldsymbol{\xi} - b_n(\boldsymbol{x}) + c_n t - \lambda \sum_{j\in[m]} \boldsymbol{\nu}_j \le q_i \quad \forall (\eta,\boldsymbol{\nu},\boldsymbol{\xi})\in\Gamma^i_{\eta,\nu,\xi}, n\in[N+1],\, i\in[M] \\ & -\boldsymbol{\nu}_i \le \boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i \le \boldsymbol{\nu}_i \quad \forall i\in[M] \\ & \lambda \ge 0, \boldsymbol{q}\in\mathbb{R}^M\,, \end{aligned}$$

where $\Gamma^i_{\eta,\nu,\xi}$ is the polyhedron capturing the joint feasible assignments for $\eta$, $\boldsymbol{\nu}$ and $\boldsymbol{\xi}$.

It is therefore easy to see that problem (8) can be reformulated as a multistage robust linear optimization problem, given that $\mathcal{X}$ and $\Xi$ are assumed polyhedral. A similar argument holds for the case of a $\ell_\infty$-norm Wasserstein distance metric. □

### A.7. Proof of Theorem 2

Our proof mainly employs the Fenchel Robust Counterpart theory in Ben-Tal et al. (2015) to derive the equivalent reformulations of robust constraints (10b) and (10c).

First, for the robust linear constraint (10b), one can directly derive the following equivalent formulation which verifies the two boundary scenarios $\bar{t}^-_k$ and $\bar{t}^+_k$, given that the constraint function is linear in $t$,

$$\lambda_k\epsilon + \frac{1}{M}\sum_{i=1}^{M} (q_{ik}\bar{t}^+_k + \bar{q}_{ik}) \le 0, \qquad \forall k\in[K] \tag{27a}$$

$$\lambda_k \epsilon + \frac{1}{M}\sum_{i=1}^{M}(q_{ik}\bar{t}_k^- + \bar{q}_{ik}) \le 0, \qquad \forall k \in [K]. \tag{27b}$$

Second, for constraint (10c), for any fixed $k \in [K], i \in [M]$ and $n \in [N+1]$, let $\phi_n^{ik}(\boldsymbol{x},\boldsymbol{\xi},t) := g_n(\boldsymbol{x},\boldsymbol{\xi},t) - \lambda_k\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i\| - \bar{q}_{ik} - q_{ik}t$. Note that, $\phi_n^{ik}(\boldsymbol{x},\boldsymbol{\xi},t)$ is convex in $\boldsymbol{x}$ for all $\boldsymbol{\xi} \in \Xi$ and $t \in \bar{\mathcal{T}}$ since it is the some of two concave functions and an affine one. Based on the Fenchel duality theory (see Ben-Tal et al. (2015)), a robust constraint that takes the form of

$$\phi(\boldsymbol{x},\boldsymbol{\xi},t) \le 0 \qquad \forall \boldsymbol{\xi} \in \Xi,\ t \in \mathcal{T}_k,$$

is equivalent to

$$\exists\, \boldsymbol{v} \in \mathbb{R}^m, u \in \mathbb{R}, \quad \delta(\boldsymbol{v}|\Xi) + \delta(u|\mathcal{T}_k) - \phi_*(\boldsymbol{x},u,\boldsymbol{v}) \le 0,$$

where $\delta(\boldsymbol{v}|\Xi)$ and $\delta(u|\mathcal{T}_k)$ are support functions, and $\phi_*(\boldsymbol{x},u,\boldsymbol{v})$ is the partial concave conjugate function of $\phi(\boldsymbol{x},\boldsymbol{\xi},t)$. More specifically, one can directly derive the following equivalent formulation of $\delta(u|\mathcal{T}_k)$ which verifies the two boundary scenarios $\bar{t}_k^-$ and $\bar{t}_k^+$, given that the constraint function is linear in $t$,

$$\delta(u|\mathcal{T}_k) = \sup_{t \in \mathbb{R}: \bar{t}_k^- \le t \le \bar{t}_k^+} u\,t \; = \max\left(u\bar{t}_k^-, u\bar{t}_k^+\right).$$

As for the partial concave conjugate of $\phi_n^{ik}(\boldsymbol{x},\boldsymbol{\xi},t)$, we exploit the representation

$$g_n(\boldsymbol{x},\boldsymbol{\xi},t) := -\boldsymbol{a}_n(\boldsymbol{x})^\top\boldsymbol{\xi} - b_n(\boldsymbol{x}) + c_n t + \inf_{\boldsymbol{\rho}\ge 0: \sum_{n'}\rho_{n'} \le 1} \sum_{n' \in [N]} \rho_{n'}\left({\boldsymbol{a}_{n'}^0}^\top \boldsymbol{\xi} + b_{n'}^0 - t\right),$$

and have that:

$$\begin{aligned}
\phi_{n*}^{ik}(\boldsymbol{x},u,\boldsymbol{v}) &= \inf_{\boldsymbol{\xi}\in\mathbb{R}^m} \boldsymbol{v}^\top\boldsymbol{\xi} + \inf_{t\in\mathbb{R}} tu - \left(g_n(\boldsymbol{x},\boldsymbol{\xi},t) - \lambda_k\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i\| - \bar{q}_{ik} - q_{ik}t\right) \\
&= \inf_{\boldsymbol{\xi}\in\mathbb{R}^m} \inf_{t\in\mathbb{R}} \boldsymbol{v}^\top\boldsymbol{\xi} + tu + \boldsymbol{a}_n(\boldsymbol{x})^\top\boldsymbol{\xi} + b_n(\boldsymbol{x}) - c_n t + \lambda_k\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i\| + \bar{q}_{ik} + q_{ik}t \\
&\qquad + \sup_{\boldsymbol{\rho}\ge 0: \sum_{n'}\rho_{n'} \le 1} \sum_{n'\in[N]} \rho_{n'}\left(-{\boldsymbol{a}_{n'}^0}^\top\boldsymbol{\xi} - b_{n'}^0 + t\right) \\
&= \sup_{\boldsymbol{\rho}\ge 0: \sum_{n'}\rho_{n'} \le 1} \inf_{\boldsymbol{\xi}\in\mathbb{R}^m} (\boldsymbol{v} + \boldsymbol{a}_n(\boldsymbol{x}) - \sum_{n'\in[N]} \rho_{n'}\boldsymbol{a}_{n'}^0)^\top\boldsymbol{\xi} + \lambda_k\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i\| + \inf_{t\in\mathbb{R}}(u + \sum_{n'\in[N]}\rho_{n'} - c_n + q_{ik})t \\
&\qquad + \bar{q}_{ik} + b_n(\boldsymbol{x}) - \sum_{n'\in[N]} \rho_{n'} b_{n'}^0 \\
&= \sup_{\boldsymbol{\rho}\ge 0: \sum_{n'}\rho_{n'} \le 1} \inf_{\boldsymbol{\xi}\in\mathbb{R}^m} \boldsymbol{w}^n(\boldsymbol{x},\boldsymbol{\rho})^\top\boldsymbol{\xi} + \lambda_k\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_i\| + \inf_{t\in\mathbb{R}}(u + q_{ik} - c_n + \sum_{n'\in[N]}\rho_{n'})t \\
&\qquad + \bar{q}_{ik} + b_n(\boldsymbol{x}) - \sum_{n'\in[N]} \rho_{n'} b_{n'}^0 \\
&= \sup_{\boldsymbol{\rho}\ge 0: \sum_{n'}\rho_{n'} \le 1} \boldsymbol{w}^n(\boldsymbol{x},\boldsymbol{\rho})^\top\hat{\boldsymbol{\xi}}_i - \lambda_k\left(\sup_{\boldsymbol{\zeta}\in\mathbb{R}^m}\left(-\frac{\boldsymbol{w}^n(\boldsymbol{x},\boldsymbol{\rho})}{\lambda_k}\right)^\top\boldsymbol{\zeta} - \|\boldsymbol{\zeta}\|\right) + \inf_{t\in\mathbb{R}}(u + q_{ik} - c_n + \sum_{n'\in[N]}\rho_{n'})t
\end{aligned}$$

$$
\begin{aligned}
&+\bar{q}_{ik} + b_n(\boldsymbol{x}) - \sum_{n' \in [N]} \rho_{n'} b^0_{n'} \\
= &\begin{cases}
\sup\limits_{\boldsymbol{\rho}, \boldsymbol{w}} \boldsymbol{w}^\top \hat{\boldsymbol{\xi}}_i + \bar{q}_{ik} + b_n(\boldsymbol{x}) - \sum\limits_{n' \in [N]} \rho_{n'} b^0_{n'} \\
\text{s.t. } \|\boldsymbol{w}\|_* \leq \lambda_k \\
\quad \boldsymbol{w} = \boldsymbol{v} + \boldsymbol{a}_n(\boldsymbol{x}) - \sum\limits_{n' \in [N]} \rho_{n'} \boldsymbol{a}^0_{n'} \\
\quad u + q_{ik} + \sum\limits_{n' \in [N]} \rho_{n'} - c_n = 0 \\
\quad \sum\limits_{n' \in [N]} \rho_{n'} \leq 1 \\
\quad \boldsymbol{\rho} \geq 0,
\end{cases}
\end{aligned}
$$

where we replaced $\boldsymbol{w}^n(\boldsymbol{x}, \boldsymbol{\rho}) := \boldsymbol{v} + \boldsymbol{a}_n(\boldsymbol{x}) - \sum\limits_{n' \in [N]} \rho_{n'} \boldsymbol{a}^0_{n'}$ and where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. In details, the first equality straightforwardly follows from the definition of conjugate function, and second equality follows from the definition of $g_n(\boldsymbol{x}, \boldsymbol{\xi}, t)$. The third equality holds by applying Sion's minimax theorem (see Sion et al. (1958)). Finally, the last equality is based on the definition of the conjugate of the norm.

We thus conclude that, for any fixed $k$, $i$, and $n$, constraint (10c) is equivalent to the existence of some $u \in \mathbb{R}$, $\boldsymbol{\rho} \in \mathbb{R}^N$, $\boldsymbol{w} \in \mathbb{R}^m$, and $\boldsymbol{v} \in \mathbb{R}^m$ such that:

$$
\begin{aligned}
&\delta(\boldsymbol{v} \mid \Xi) + u\bar{t}^-_k - \boldsymbol{w}^\top \hat{\boldsymbol{\xi}}_i - \bar{q}_{ik} - b_n(\boldsymbol{x}) + \sum_{n' \in [N]} \rho_{n'} b^0_{n'} \leq 0 && \text{(28a)} \\
&\delta(\boldsymbol{v} \mid \Xi) + u\bar{t}^+_k - \boldsymbol{w}^\top \hat{\boldsymbol{\xi}}_i - \bar{q}_{ik} - b_n(\boldsymbol{x}) + \sum_{n' \in [N]} \rho_{n'} b^0_{n'} \leq 0 && \text{(28b)} \\
&\boldsymbol{w} = \boldsymbol{v} + \boldsymbol{a}_n(\boldsymbol{x}) - \sum_{n' \in [N]} \rho_{n'} \boldsymbol{a}^0_{n'} && \text{(28c)} \\
&\|\boldsymbol{w}\|_* \leq \lambda_k && \text{(28d)} \\
&u + q_{ik} + \sum_{n' \in [N]} \rho_{n'} - c_n = 0 && \text{(28e)} \\
&\sum_{n' \in [N]} \rho_{n'} \leq 1 && \text{(28f)} \\
&\boldsymbol{\rho} \geq 0. && \text{(28g)}
\end{aligned}
$$

This completes our reformulation that is presented in the theorem.

Moreover, under Assumption 3 and when $\Xi$ is polyhedral, one has that $\|\boldsymbol{w}\|_*$, $\delta(\boldsymbol{v} \mid \Xi)$, and $\mathcal{X}$ are LP representable, hence the problem can be reformulated as a linear program. □

### A.8. Proof of Proposition 6

First, one concludes that the optimal value of problem (11) provides a lower bound for the optimal value of (8) based on the fact that the latter is obtained by relaxing constraint (8b).

Given a finite scenario set $\hat{\mathcal{T}} = \{\hat{t}_1, \cdots, \hat{t}_k, \cdots, \hat{t}_K\}$, to obtain a finite-dimensional representation of problem (11), we just need to derive the equivalent reformulation of robust constraint (11c). For this purpose, we can apply exactly the same steps as used in the proof of Theorem 2, while letting $\bar{t}_k^- = \bar{t}_k^+ = \hat{t}_k$, $\bar{q}_{ik} = q_{ik}$ and $q_{ik} = 0$. We obtain that for any fixed $i$, $n$, and $k$, constraint (11c) is equivalent to the condition that there exists $\boldsymbol{\rho} \in \mathbb{R}^N$, $\boldsymbol{w} \in \mathbb{R}^m$, and $\boldsymbol{v} \in \mathbb{R}^m$ such that:

$$\begin{aligned}
&\delta(\boldsymbol{w} + \sum_{n' \in [N]} \rho_{n'} \boldsymbol{a}_n^0 - \boldsymbol{a}_{n'}(\boldsymbol{x}) \mid \Xi) - \boldsymbol{w}^\top \hat{\boldsymbol{\xi}}_i + \sum_{n' \in [N]} \rho_{n'} b_{n'}^0 \\
&- b_n(\boldsymbol{x}) - (\sum_{n' \in [N]} \rho_{n'} - c_n)\hat{t}_k - q_{ik} \leq 0 \\
&\|\boldsymbol{w}\|_* \leq \lambda_k \\
&\sum_{n' \in [N]} \rho_{n'} \leq 1 \\
&\boldsymbol{\rho} \geq 0.
\end{aligned}$$

Finally, we obtain the finite-dimensional convex optimization formulation of problem (11). Moreover, under Assumption 3 and when $\Xi$ is polyhedral, one has that $\|\boldsymbol{w}\|_*$, $\delta(\cdot \mid \Xi)$, and $\mathcal{X}$ are LP representable, hence the problem can be reformulated as a linear program. □

### A.9. Proof of Proposition 7

The first part of our proof can be straightforward given from the way we construct the finite scenarios set that is described in Algorithm 1, namely, $\hat{\mathcal{T}}^\ell := \hat{\mathcal{T}}^{\ell-1} \bigcup \hat{\mathcal{A}}^\ell$, which means that $\hat{\mathcal{T}}^{\ell-1}$ is a subset of $\hat{\mathcal{T}}^\ell$. Thus, the lower bounding problem (11) with $\hat{\mathcal{T}}^\ell$ is a relaxation to problem (11) with $\hat{\mathcal{T}}^{\ell+1}$, and we have $\mathrm{LB}^{\ell+1} \geq \mathrm{LB}^\ell$.

An analogous argument can be made about how $\mathrm{UB}^\ell$ is generated. Indeed, problem (10) under $\mathscr{P}^{\ell+1}$ can be considered a relaxation of problem (10) under $\mathscr{P}^\ell$ since any optimal solution $(\boldsymbol{x}^\ell, \boldsymbol{\lambda}^\ell, \boldsymbol{q}^\ell, \bar{\boldsymbol{q}}^\ell)$ of the latter can be used to produce a feasible solution $(\boldsymbol{x}^{\ell+1}, \boldsymbol{\lambda}^{\ell+1}, \boldsymbol{q}^{\ell+1}, \bar{\boldsymbol{q}}^{\ell+1})$ to the former by using the following construction:

$$\begin{aligned}
&\boldsymbol{x}^{\ell+1} := \boldsymbol{x}^\ell, && \boldsymbol{\lambda}_k^{\ell+1} := \sum_{k' \in [|\mathscr{P}^\ell|]: \mathcal{T}_k^{\ell+1} \subseteq \mathcal{T}_{k'}^\ell} \boldsymbol{\lambda}_{k'}^\ell, \ \forall k \in [|\mathscr{P}^{\ell+1}|], \\
&\boldsymbol{q}_{ik}^{\ell+1} := \sum_{k' \in [|\mathscr{P}^\ell|]: \mathcal{T}_k^{\ell+1} \subseteq \mathcal{T}_{k'}^\ell} \boldsymbol{q}_{ik'}^\ell, && \bar{\boldsymbol{q}}_{ik}^{\ell+1} := \sum_{k' \in [|\mathscr{P}^\ell|]: \mathcal{T}_k^{\ell+1} \subseteq \mathcal{T}_{k'}^\ell} \bar{\boldsymbol{q}}_{ik'}^\ell, \ \forall k \in [|\mathscr{P}^{\ell+1}|].
\end{aligned}$$

Thus, for the conservative approximation model (10), we can easily obtain $\mathrm{UB}^\ell \geq \mathrm{UB}^{\ell+1}$. □

### A.10. Proof of Proposition 8

Our proof follows similar steps as in the proof of Proposition 5. Based on the result in Lemma 1, we first rewrite D-DRSSDCP (14) with $\mathcal{P}_{\mathrm{W}^2}^1$ as follows:

$$\underset{\boldsymbol{x} \in \mathcal{X}}{\text{minimize}} \ \boldsymbol{c}^\top \boldsymbol{x} \tag{29a}$$

$$\text{subject to} \sup_{\mathbb{P}\in\mathcal{P}^1_{\mathrm{W}^2}} \mathbb{E}_{\mathbb{P}}\left[(t-f(\boldsymbol{x},\boldsymbol{\xi}))^+ - (t-f_0(\boldsymbol{\zeta}))^+\right] \leq 0 \qquad \forall t\in\mathbb{R}. \tag{29b}$$

Similar to Lemma 4, here we also derive an equivalent representation of constraint (29b) by restricting $t\in\bar{\mathcal{T}}' := [t'_{min}, t'_{max}]$, where $t'_{min} := \inf_{\boldsymbol{\zeta}\in\Xi_\zeta} f_0(\boldsymbol{\zeta})$ and $t'_{max} := \sup_{\boldsymbol{\zeta}\in\Xi_\zeta} f_0(\boldsymbol{\zeta})$. Under the definition of $\mathcal{P}^1_{\mathrm{W}^2}$, we can further rewrite constraint (29b) as

$$\sup_{\mathbb{P}_\xi\in\mathcal{P}^1_\xi(\hat{\mathbb{P}}_\xi,\epsilon_\xi)} \mathbb{E}_{\mathbb{P}_\xi}\left[(t-f(\boldsymbol{x},\boldsymbol{\xi}))^+\right] + \sup_{\mathbb{P}_\zeta\in\mathcal{P}^1_\zeta(\hat{\mathbb{P}}_\zeta,\epsilon_\zeta)} \mathbb{E}_{\mathbb{P}_\zeta}\left[-(t-f_0(\boldsymbol{\zeta}))^+\right] \leq 0 \qquad \forall t\in\bar{\mathcal{T}}'. \tag{30}$$

Under assumptions 2 and 4, by using the strong duality results for Wasserstein DRO in Mohajerin Esfahani and Kuhn (2018), for any given fixed $t\in\bar{\mathcal{T}}'$, we can derive the equivalent reformulation for constraint (30). Then we can easily obtain the multistage robust optimization formulation that is presented in the proposition, which also takes the form of $\min_{\boldsymbol{x}}\text{-}\sup_t\text{-}\min_{\lambda^1,\lambda^2,\boldsymbol{q},\boldsymbol{r}}\text{-}\sup_{\boldsymbol{\xi},\boldsymbol{\zeta}}$. □

### A.11. Proof of Proposition 9

Our proof reuses some of the steps presented in the proof of Proposition 11 as we will employ Proposition 3.2 in Dentcheva and Ruszczynski (2003). Yet, the type-1 Wasserstein DRO reformulations from Mohajerin Esfahani and Kuhn (2018) are now necessary. First, when $\epsilon_\zeta = 0$, $\mathcal{P}^1_\zeta(\hat{\mathbb{P}}_\zeta,\epsilon_\zeta)$ reduces to a singleton $\{\hat{\mathbb{P}}_\zeta\}$. DRSSD constraint (30) can be rewritten as

$$\sup_{\mathbb{P}_\xi\in\mathcal{P}^1_\xi(\hat{\mathbb{P}}_\xi,\epsilon_\xi)} \mathbb{E}_{\mathbb{P}_\xi}\left[(t-f(\boldsymbol{x},\boldsymbol{\xi}))^+\right] \leq \frac{1}{M_\zeta}\sum_{i\in[M_\zeta]}\left(t-f_0(\hat{\boldsymbol{\zeta}}_i)\right)^+ \qquad \forall t\in\bar{\mathcal{T}}'. \tag{31}$$

Since $\mathbb{P}_\zeta$ is supported on $\{\hat{\boldsymbol{\zeta}}_i\}_{i=1}^{M_\zeta}$, and $f_0(\boldsymbol{\zeta})$ is independent of decision $\boldsymbol{x}$, so $f_0(\boldsymbol{\zeta}_i)$ is always supported on $\{f_0(\hat{\boldsymbol{\zeta}}_i)\}_{i=1}^{M_\zeta}$. Thus, for any given $\mathbb{P}_\xi\in\mathcal{P}^1_\xi(\hat{\mathbb{P}}_\xi,\epsilon_\xi)$, one can employ Proposition 3.2 in Dentcheva and Ruszczynski (2003) to equivalently replace $\bar{\mathcal{T}}'$ with $\{t_1,\cdots,t_{M_\zeta}\}$, where each $t_j := f_0(\hat{\boldsymbol{\zeta}}_j)$. Let us define $\gamma_j = \frac{1}{M_\zeta}\sum_{i\in[M_\zeta]}\left(t_j - f_0(\hat{\boldsymbol{\zeta}}_i)\right)^+$ for all $j\in[M_\zeta]$, then constraint (31) can be further rewritten as the following finite number of constraints,

$$\sup_{\mathbb{P}_\xi\in\mathcal{P}^1_\xi(\hat{\mathbb{P}}_\xi,\epsilon_\xi)} \mathbb{E}_{\mathbb{P}_\xi}\left[(t_j-f(\boldsymbol{x},\boldsymbol{\xi}))^+\right] \leq \gamma_j \qquad \forall j\in[M_\zeta]. \tag{32}$$

Based on assumptions 2 and 4, following the similar steps in Mohajerin Esfahani and Kuhn (2018), we can derive the equivalent condition that there exists $\boldsymbol{\lambda}\in\mathbb{R}^{M_\zeta}$ such that:

$$\lambda_j\epsilon_\xi + \frac{1}{M_\xi}\sum_{i\in[M_\xi]}\sup_{\boldsymbol{\xi}\in\Xi_\xi}\left((t_j-f(\boldsymbol{x},\boldsymbol{\xi}))^+ - \lambda_j\|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_i\|\right) \leq \gamma_j \qquad \forall j\in[M_\zeta]. \tag{33}$$

Based on Assumption 2, $(t_j - f(\boldsymbol{x},\boldsymbol{\xi}))^+ = \max_{n\in[N+1]} c_n t_j - \boldsymbol{a}_n(\boldsymbol{x})^\top\boldsymbol{\xi} - b_n(\boldsymbol{x})$, where $a_{N+1} = b_{N+1} = c_{N+1} = 0$ and $c_n = 1$ for all $n\in[N]$. By following the Fenchel Robust Counterpart theory in Ben-Tal et al. (2015), constraint (33) is equivalent to the condition that there exists $\boldsymbol{\lambda}\in\mathbb{R}^{M_\zeta}$, $\boldsymbol{s}\in\mathbb{R}^{M_\xi\times M_\zeta}$, $\boldsymbol{v}\in\mathbb{R}^{M_\xi\times M_\zeta\times(N+1)\times m}$, and $\boldsymbol{w}\in\mathbb{R}^{M_\xi\times M_\zeta\times(N+1)\times m}$ such that:

$$\lambda_j \epsilon_\xi + \frac{1}{M_\xi} \sum_{i \in [M_\xi]} s_{ij} \leq \gamma_j \qquad \forall j \in [M_\zeta] \tag{34}$$

$$\delta(\boldsymbol{v}_{ijn} | \Xi_\xi) - \boldsymbol{w}_{ijn}^\top \hat{\boldsymbol{\xi}}_i - b_n(\boldsymbol{x}) + c_n t_j \leq s_{ij} \qquad \forall i \in [M_\xi], j \in [M_\zeta], n \in [N+1] \tag{35}$$

$$\|\boldsymbol{w}_{ijn}\|_* \leq \lambda_j \qquad \forall i \in [M_\xi], j \in [M_\zeta], n \in [N+1] \tag{36}$$

$$\boldsymbol{w}_{ijn} = \boldsymbol{v}_{ijn} + \boldsymbol{a}_n(\boldsymbol{x}) \qquad \forall i \in [M_\xi], j \in [M_\zeta], n \in [N+1] \tag{37}$$

$$\boldsymbol{\lambda} \geq 0. \tag{38}$$

Thus we finally derive the finite-dimensional convex optimization problem (17). Moreover, under Assumption 3 and when $\Xi_\xi$ is polyhedral, one has that $\|\boldsymbol{w}\|_*$, $\delta(\boldsymbol{v} \mid \Xi_\xi)$, and $\mathcal{X}$ are LP representable, hence problem (17) can be reformulated as a linear program. □

## Appendix B: Supplementary Material

### B.1. Alternative Representations in Montes et al. (2014)

In this appendix, we attempt to make a comparison to the alternative representations that have been proposed in Montes et al. (2014) for extending the notion of stochastic dominance to an ambiguous probability space. In their Definition 5, the authors propose the following six different formulations:

$$X \succeq^{(39)} Y \Leftrightarrow F_X^{\mathbb{P}_1} \succeq_{(k)} F_Y^{\mathbb{P}_2},\ \forall \mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P} \tag{39}$$

$$X \succeq^{(40)} Y \Leftrightarrow \exists \mathbb{P}_1 \in \mathcal{P},\ F_X^{\mathbb{P}_1} \succeq_{(k)} F_Y^{\mathbb{P}_2},\ \forall \mathbb{P}_2 \in \mathcal{P} \tag{40}$$

$$X \succeq^{(41)} Y \Leftrightarrow \forall \mathbb{P}_2 \in \mathcal{P},\ \exists \mathbb{P}_1 \in \mathcal{P},\ F_X^{\mathbb{P}_1} \succeq_{(k)} F_Y^{\mathbb{P}_2} \tag{41}$$

$$X \succeq^{(42)} Y \Leftrightarrow \exists \mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P},\ F_X^{\mathbb{P}_1} \succeq_{(k)} F_Y^{\mathbb{P}_2} \tag{42}$$

$$X \succeq^{(43)} Y \Leftrightarrow \exists \mathbb{P}_2 \in \mathcal{P},\ F_X^{\mathbb{P}_1} \succeq_{(k)} F_Y^{\mathbb{P}_2},\ \forall \mathbb{P}_1 \in \mathcal{P} \tag{43}$$

$$X \succeq^{(44)} Y \Leftrightarrow \forall \mathbb{P}_1 \in \mathcal{P},\ \exists \mathbb{P}_2 \in \mathcal{P},\ F_X^{\mathbb{P}_1} \succeq_{(k)} F_Y^{\mathbb{P}_2}. \tag{44}$$

Yet, we can easily observe that (39), (40), and (43) are not necessarily reflexive while (42) is not necessarily transitive. Specifically, take the example of a random variable $X$ with an ambiguity set $\mathcal{P} = \{\mathbb{P}_1, \mathbb{P}_2\}$ such that neither $F_X^{\mathbb{P}_1} \not\succeq_{(k)} F_X^{\mathbb{P}_2}$ nor $F_X^{\mathbb{P}_2} \not\succeq_{(k)} F_X^{\mathbb{P}_1}$: e.g. in the case of $k = 1$, when $F_X^{\mathbb{P}_1}$ is Bernoulli $\{0, 1\}$ with $\mathbb{P}_1(X = 1) = 50\%$ while $F_X^{\mathbb{P}_2}$ is a guaranteed return of 0.6. In this case we have that

$$F_X^{\mathbb{P}_1} \not\succeq_{(k)} F_X^{\mathbb{P}_2} \Rightarrow \exists \mathbb{P}_1, \mathbb{P}_2,\ F_X^{\mathbb{P}_1} \not\succeq_{(k)} F_X^{\mathbb{P}_2} \Rightarrow X \not\succeq^{(39)} X$$

$$F_X^{\mathbb{P}_1} \not\succeq_{(k)} F_X^{\mathbb{P}_2} \,\&\, F_X^{\mathbb{P}_2} \not\succeq_{(k)} F_X^{\mathbb{P}_1} \Rightarrow \nexists \mathbb{P}_1, \forall \mathbb{P}_2,\ F_X^{\mathbb{P}_1} \succeq_{(k)} F_X^{\mathbb{P}_2} \Rightarrow X \not\succeq^{(40)} X$$

$$F_X^{\mathbb{P}_1} \not\succeq_{(k)} F_X^{\mathbb{P}_2} \,\&\, F_X^{\mathbb{P}_2} \not\succeq_{(k)} F_X^{\mathbb{P}_1} \Rightarrow \nexists \mathbb{P}_2, \forall \mathbb{P}_1,\ F_X^{\mathbb{P}_1} \succeq_{(k)} F_X^{\mathbb{P}_2} \Rightarrow X \not\succeq^{(43)} X.$$

To demonstrate that $\succeq^{(42)}$ is non-transitive, we consider the following three random variables which are deterministic under both $\mathbb{P}_1$ and $\mathbb{P}_2$:

$$F_A^{\mathbb{P}}(y) := 1\{y \geq 2\} \qquad F_B^{\mathbb{P}}(y) := \begin{cases} 1\{y \geq 0\} & \text{if } \mathbb{P} = \mathbb{P}_1 \\ 1\{y \geq 4\} & \text{if } \mathbb{P} = \mathbb{P}_2 \end{cases} \qquad F_C^{\mathbb{P}}(y) := 1\{y \geq 1\}.$$

Thus, we get:

$$F_B^{\mathbb{P}_2} \succeq_{(k)} F_A^{\mathbb{P}_2} \Rightarrow B \succeq^{(42)} A \qquad \& \qquad F_C^{\mathbb{P}_1} \succeq_{(k)} F_B^{\mathbb{P}_1} \Rightarrow C \succeq^{(42)} B.$$

Yet,

$$\forall \mathbb{P}, \mathbb{Q} \in \{\mathbb{P}_1, \mathbb{P}_2\},\ F_C^{\mathbb{P}} = 1\{y \geq 1\} \not\succeq_{(k)} 1\{y \geq 2\} = F_A^{\mathbb{Q}} \Rightarrow C \not\succeq^{(42)} A.$$

Looking now more closely at (41) and (44), one can confirm that the two extensions satisfy ambiguity monotonicity, yet we have that:

$$X \succeq^{(3)} Y \Leftrightarrow F_X^{\mathbb{P}} \succeq_{(k)} F_Y^{\mathbb{P}},\ \forall \mathbb{P} \in \mathcal{P} \Rightarrow \forall \mathbb{P}_2 \in \mathcal{P},\ \exists \mathbb{P}_1 \in \mathcal{P},\ F_X^{\mathbb{P}_1} \succeq_{(k)} F_Y^{\mathbb{P}_2} \Rightarrow X \succeq^{(41)} Y$$

and

$$X \succeq^{(3)} Y \Leftrightarrow F_X^{\mathbb{P}} \succeq F_Y^{\mathbb{P}},\ \forall \mathbb{P} \in \mathcal{P} \Rightarrow \forall \mathbb{P}_1 \in \mathcal{P},\ \exists \mathbb{P}_2 \in \mathcal{P},\ F_X^{\mathbb{P}_1} \succeq_{(k)} F_Y^{\mathbb{P}_2} \Rightarrow X \succeq^{(44)} Y\,.$$

This allows us to conclude that the extension of stochastic dominance that we propose in Section 4 is more indecisive than the formulations in (41) and (44). In other words, our formulation makes less assumptions on how the ambiguity about stochastic dominance is resolved.

### B.2. DRSSDCP with a type-infinity Wasserstein Ambiguity Set

In this appendix, we will show that the proposed solution scheme can be straightforwardly extended for DRSSDCP with a type-infinity Wasserstein ambiguity set. In the following, we present how to derive its multistage robust optimization reformulation under mild conditions.

Extending Definition 3 to the case of $r = \infty$, the Wasserstein metric between distributions $\mathbb{P}_1 \in \mathcal{M}(\Xi)$ and $\mathbb{P}_2 \in \mathcal{M}(\Xi)$ can be defined as

$$d_{\mathrm{W}}^{\infty}(\mathbb{P}_1, \mathbb{P}_2) = \inf_{\mathbb{Q} \in \mathcal{M}(\mathbb{P}_1, \mathbb{P}_2)} \left\{ \operatorname*{ess\,sup}_{\mathbb{Q}} \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\| \right\},$$

where $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \sim \mathbb{Q}$.

The following proposition gives the multistage robust reformulation of DRSSDCP with $\mathcal{P}_{\mathrm{W}}^{\infty}(\hat{\mathbb{P}}, \epsilon)$ under mild conditions, which takes the similar problem structure with problem (8). Hence, we can directly adapt the solution scheme that is proposed in sections 5, 6 and 7 to solve problem (45).

Proposition 10. *Under assumptions 1, 2 and $\mathcal{P}_W^{\infty}(\hat{\mathbb{P}},\epsilon)$ with $\epsilon \in (0,\infty)$, DRSSDCP* (6) *coincides with the optimal value of the following multistage robust optimization problem:*

$$\begin{aligned}
&\underset{\boldsymbol{x}\in\mathcal{X}}{\text{minimize}}\ \boldsymbol{c}^\top \boldsymbol{x} && && \text{(45a)}\\
&\text{subject to } L(\boldsymbol{x},t) \leq 0 && \forall t \in \bar{\mathcal{T}}, && \text{(45b)}
\end{aligned}$$

*where $\bar{\mathcal{T}} := [t_{min}, t_{max}]$ with $t_{min} := \inf_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{\xi})$ and $t_{max} := \sup_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{\xi})$, and where*

$$\begin{aligned}
L(\boldsymbol{x},t) := \quad & \inf_{\boldsymbol{\lambda},\boldsymbol{q}}\ \frac{1}{M}\sum_{i\in[M]} q_i && && \text{(46a)}\\
& \text{subject to}\quad \lambda_i\epsilon + g(\boldsymbol{x},\boldsymbol{\xi},t) - \lambda_i\|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_i\| \leq q_i && \forall \boldsymbol{\xi}\in\Xi, i\in[M] && \text{(46b)}\\
& \boldsymbol{\lambda}\geq 0, \boldsymbol{q}\in\mathbb{R}^M, && && \text{(46c)}
\end{aligned}$$

*where $g(\boldsymbol{x},\boldsymbol{\xi},t) = (t - f(\boldsymbol{x},\boldsymbol{\xi}))^+ - (t - f_0(\boldsymbol{\xi}))^+$. Moreover, it can be reformulated as a multistage robust linear optimization problem when $\ell_1$-norm (or $\ell_\infty$-norm) of the Wasserstein distance metric is used, and when $\mathcal{X}$, $\Xi$ are polyhedral.*

. **Proof of Proposition 10** Our proof follows the similar steps of the proof for Proposition 5 in Appendix A.6. For the sake of completeness and simplicity, here we mention the key steps of the proof.

Similar to Lemma 4, under assumptions 1 and 2, constraint (6b) is equivalent to

$$\sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{W}}^{\infty}(\hat{\mathbb{P}},\epsilon)} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x},\boldsymbol{\xi},t)\right] \leq 0 \quad \forall t \in \bar{\mathcal{T}},$$

where $\bar{\mathcal{T}} := [t_{min}, t_{max}]$ with $t_{min} := \inf_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{\xi})$ and $t_{max} := \sup_{\boldsymbol{\xi}\in\Xi} f_0(\boldsymbol{\xi})$.

Based on Lemma 5, we know that $g(\boldsymbol{x},\boldsymbol{\xi},t)$ is a maximum of piecewise linear concave functions in $\boldsymbol{\xi}$ and $t$ for all $\boldsymbol{x}\in\mathbb{R}^m$ if Assumption 2 holds. By following the similar steps in Bertsimas et al. (2021), for any given fixed $t\in\bar{\mathcal{T}}$, the worst-case expectation $\sup_{\mathbb{P}\in\mathcal{P}_{\mathrm{W}}^{\infty}(\hat{\mathbb{P}},\epsilon)} \mathbb{E}_{\mathbb{P}}\left[g(\boldsymbol{x},\boldsymbol{\xi},t)\right]$ coincides with the optimal value of the following optimization problem:

$$\begin{aligned}
& \inf_{\boldsymbol{\lambda},\boldsymbol{q}}\ \frac{1}{M}\sum_{i\in[M]} q_i && && \text{(47a)}\\
& \text{subject to}\quad \lambda_i\epsilon + g(\boldsymbol{x},\boldsymbol{\xi},t) - \lambda_i\|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_i\| \leq q_i && \forall \boldsymbol{\xi}\in\Xi, i\in[M] && \text{(47b)}\\
& \boldsymbol{\lambda}\geq 0, \boldsymbol{q}\in\mathbb{R}^M. && && \text{(47c)}
\end{aligned}$$

Therefore, DRSSDCP (4) with $\mathcal{P}_{\mathrm{W}}^{\infty}(\hat{\mathbb{P}},\epsilon)$ is equivalent to the multistage robust optimization problem (45). Moreover, it is easy to obtain that problem (45) can be reformulated as a multistage linear optimization problem if either $\ell_1$-norm or $\ell_\infty$-norm in constraint (47b) is used, and if $\mathcal{X}$ and $\Xi$ are polyhedral. □

## B.3. Robust Second-order Stochastic Dominance Constrained Portfolio Optimization Problem in Sehgal and Mehra (2020)

In this appendix, we briefly show that our D-DRSSDCP with $\epsilon_\xi > \epsilon_\zeta = 0$ can recover the robust second-order stochastic dominance constrained portfolio optimization problem in Sehgal and Mehra (2020), which can be summarized in the following proposition.

Proposition 11. *The robust second-order stochastic dominance constraint proposed in Sehgal and Mehra (2020) is a special case of D-DRSSDCP with $\mathcal{P}^{\infty}_{\mathrm{W}^2}$, $\epsilon_\zeta = 0$, $\Xi_\xi = \mathbb{R}^m$, and the semimetric*

$$d(\boldsymbol{\xi}, \boldsymbol{\xi}_0) := \begin{cases} \infty & \text{if } \|\boldsymbol{\xi} - \boldsymbol{\xi}_0\|_\infty > 1 \\ \|\boldsymbol{\xi} - \boldsymbol{\xi}_0\|_1 & \text{otherwise.} \end{cases}$$

. **Proof of Proposition 11** Our proof follows similar steps as presented in the proof of Proposition 3 in Bertsimas et al. (2021) and the well-known duality results for the budgeted uncertainty set. First, let us remind the reader that the type-$\infty$ Wasserstein distance is defined as $d^{\infty}_{\mathrm{W}}(\mathbb{P}_1, \mathbb{P}_2) = \inf_{\mathbb{Q} \in \mathcal{M}(\mathbb{P}_1, \mathbb{P}_2)} \mathbb{Q}\text{-}\operatorname{ess\,sup} d(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$, which implies that:

$$\begin{aligned} d^{\infty}_{\mathrm{W}}(\mathbb{P}_\xi, \hat{\mathbb{P}}_\xi) &= \inf_{\{\mathbb{Q}_i\}_{i=1}^{M_\xi} \in \mathcal{M}^{M_\xi}(\Xi_\xi) : \mathbb{P}_\xi = (1/M_\xi)\sum_i \mathbb{Q}_i} \max_{i \in [M_\xi]} \operatorname{ess\,sup}_{\mathbb{Q}_i} d(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_i) \\ &= \inf\{s : \exists \{\mathbb{Q}_i\}_{i=1}^{M_\xi} \in \mathcal{M}^{M_\xi}(\Xi_\xi) : \mathbb{P}_\xi = (1/M_\xi)\sum_i \mathbb{Q}_i, \operatorname{ess\,sup}_{\mathbb{Q}_i} d(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_i) \le s \ \forall i \in [M_\xi]\}, \end{aligned}$$

where $\operatorname{ess\,sup}_{\mathbb{Q}_i} d(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_i)$ considers $\boldsymbol{\xi} \sim \mathbb{Q}_i$. Hence, we can exploit the fact that

$$\mathcal{P}^{\infty}_{\xi}(\hat{\mathbb{P}}_\xi, \epsilon_\xi) = \{\mathbb{P}_\xi \in \mathcal{M}(\Xi_\xi) | \exists \{\mathbb{Q}_i\}_{i=1}^{M_\xi} \in \mathcal{M}^{M_\xi}(\Xi_\xi), \operatorname{ess\,sup}_{\mathbb{Q}_i} d(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_i) \le \epsilon_\xi, \mathbb{P}_\xi = (1/M_\xi)\sum_i \mathbb{Q}_i\}. \tag{48}$$

We can thus show that constraint (14b) reduces to

$$\begin{aligned} & f(\boldsymbol{x}, \boldsymbol{\xi}) \succeq^{\mathbb{P}}_{(2)} f_0(\boldsymbol{\zeta}), \ \forall \mathbb{P} \in \mathcal{P}^{\infty}_{\mathrm{W}^2} \\ & \Leftrightarrow \mathbb{E}_{\mathbb{P}_\xi}[(t - f(\boldsymbol{x}, \boldsymbol{\xi}))^+] \le \mathbb{E}_{\hat{\mathbb{P}}_\zeta}[(t - f_0(\boldsymbol{\zeta}))^+], \ \forall t \in \mathbb{R}, \ \forall \mathbb{P}_\xi \in \mathcal{P}^{\infty}_{\xi}(\hat{\mathbb{P}}_\xi, \epsilon_\xi) && (49) \\ & \Leftrightarrow \mathbb{E}_{\mathbb{P}_\xi}[(t_j - f(\boldsymbol{x}, \boldsymbol{\xi}))^+] \le \gamma_j, \ \forall j \in [M_\zeta], \ \forall \mathbb{P}_\xi \in \mathcal{P}^{\infty}_{\xi}(\hat{\mathbb{P}}_\xi, \epsilon_\xi) && (50) \\ & \Leftrightarrow \sup_{\mathbb{P}_\xi \in \mathcal{P}^{\infty}_{\xi}(\hat{\mathbb{P}}_\xi, \epsilon_\xi)} \mathbb{E}_{\mathbb{P}_\xi}[(t_j - f(\boldsymbol{x}, \boldsymbol{\xi}))^+] \le \gamma_j, \ \forall j \in [M_\zeta] && (51) \\ & \Leftrightarrow \sup_{\{\mathbb{Q}_i\}_{i=1}^{M_\xi} \in \mathcal{M}^{M_\xi}(\Xi_\xi) : \operatorname{ess\,sup}_{\mathbb{Q}_i} d(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_i) \le \epsilon_\xi} \frac{1}{M_\xi} \sum_{i=1}^{M_\xi} \mathbb{E}_{\mathbb{Q}_i}[(t_j - f(\boldsymbol{x}, \boldsymbol{\xi}))^+] \le \gamma_j, \ \forall j \in [M_\zeta] && (52) \\ & \Leftrightarrow \frac{1}{M_\xi} \sum_{i=1}^{M_\xi} \sup_{\mathbb{Q}_i \in \mathcal{M}(\Xi_\xi) : \operatorname{ess\,sup}_{\mathbb{Q}_i} d(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_i) \le \epsilon_\xi} \mathbb{E}_{\mathbb{Q}_i}[(t_j - f(\boldsymbol{x}, \boldsymbol{\xi}))^+] \le \gamma_j, \ \forall j \in [M_\zeta] && (53) \\ & \Leftrightarrow \frac{1}{M_\xi} \sum_{i=1}^{M_\xi} \sup_{\boldsymbol{\xi} \in \Xi_\xi : d(\boldsymbol{\xi}, \hat{\boldsymbol{\xi}}_i) \le \epsilon_\xi} (t_j - f(\boldsymbol{x}, \boldsymbol{\xi}))^+ \le \gamma_j, \ \forall j \in [M_\zeta] && (54) \end{aligned}$$

$$\Leftrightarrow \frac{1}{M_\xi}\sum_{i=1}^{M_\xi} \sup_{\boldsymbol{\xi}\in\Xi_\xi\cap\text{Budget}(\hat{\boldsymbol{\xi}}_i,\epsilon_\xi)} (t_j - f(\boldsymbol{x},\boldsymbol{\xi}))^+ \le \gamma_j,\, \forall j\in[M_\zeta] \tag{55}$$

$$\Leftrightarrow \exists\, \boldsymbol{d}\in\mathbb{R}_+^{M_\xi\times M_\zeta},\, \begin{cases} \frac{1}{M_\xi}\sum_{i=1}^{M_\xi} d_{ij}\le\gamma_j,\, \forall j\in[M_\zeta], \\ d_{ij}\ge \sup_{\boldsymbol{\xi}\in\Xi_\xi\cap\text{Budget}(\hat{\boldsymbol{\xi}}_i,\epsilon_\xi)} t_j - f(\boldsymbol{x},\boldsymbol{\xi}),\, \forall i\in[M_\xi], j\in[M_\zeta], \end{cases} \tag{56}$$

where $t_j := f_0(\hat{\boldsymbol{\zeta}}_j)$ and $\gamma_j := \mathbb{E}_{\hat{\mathbb{P}}_\zeta}[(t_j - f_0(\boldsymbol{\zeta}))^+]$, and where $\text{Budget}(\hat{\boldsymbol{\xi}}_i,\epsilon_\xi) := \{\boldsymbol{\xi}\mid \|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_i\|_\infty\le 1,\, \|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_i\|_1\le\epsilon_\xi\}$. In details, Equation (49) follows from the fact that $\epsilon_\zeta = 0$. Equation (50) follows from the fact that $\hat{\mathbb{P}}_\zeta$ is discrete (see Proposition 3.2 in Dentcheva and Ruszczynski (2003)). Then, we exploit equation (48) in (52) and the fact that the supremums in (53) are achieved using Dirac distributions. Finally, in (56) we employ an epigraph representation of the condition. Given that the portfolio selection problem in Sehgal and Mehra (2020) has $f(\boldsymbol{x},\boldsymbol{\xi}) := \boldsymbol{\xi}^\top\boldsymbol{x}$ and $\Xi_\xi = \mathbb{R}^m$, the rest follows from well-known reformulations of robust linear constraints with the budgeted uncertainty set (see Bertsimas and Sim (2004)). □

### B.4. Deriving Reformulation of Conservative Approximation for Problem (15)

In this appendix, we show how to derive the reformulation for the conservative approximation of problem (15) when a partition of $\bar{\mathcal{T}}$ is given.

Given the partition $\mathscr{P} := \{\mathcal{T}_k\}_{k=1}^K$, similar to Section 5.4, here we apply the piecewise constant policies to $\lambda^1(\cdot)$ and $\lambda^2(\cdot)$, and piecewise linear policies to $\boldsymbol{q}(\cdot)$, $\boldsymbol{r}(\cdot)$ respectively, namely, $\boldsymbol{\lambda}^1(t) = \sum_{k\in[K]}\lambda_k^1\mathbf{1}\{t\in\mathcal{T}_k\}$ and $\boldsymbol{q}_i(t) = \sum_{k\in[K]}(\bar{q}_{ik}+q_{ik}t)\mathbf{1}\{t\in\mathcal{T}_k\}$ for all $i\in[M_\xi]$, $\boldsymbol{\lambda}^2(t) = \sum_{k\in[K]}\lambda_k^2\mathbf{1}\{t\in\mathcal{T}_k\}$ and $\boldsymbol{r}_{i'}(t) = \sum_{k\in[K]}(\bar{r}_{i'k}+r_{i'k}t)\mathbf{1}\{t\in\mathcal{T}_k\}$ for all $i'\in[M_\zeta]$. In doing so, problem (15) can be conservatively approximated as the following optimization problem:

$$\begin{aligned}
&\underset{\boldsymbol{x},\boldsymbol{\lambda^1},\boldsymbol{\lambda^2},\boldsymbol{q},\bar{\boldsymbol{q}},\boldsymbol{r},\bar{\boldsymbol{r}}}{\text{minimize}}\ \boldsymbol{c}^\top\boldsymbol{x} && \text{(57a)}\\
&\text{subject to}\ \ \lambda_k^1\epsilon_\xi + \lambda_k^2\epsilon_\zeta + \frac{1}{M_\xi}\sum_{i\in[M_\xi]}(q_{ik}+\bar{q}_{ik}t) \\
&\qquad\qquad + \frac{1}{M_\zeta}\sum_{i'\in[M_\zeta]}(r_{i'k}+\bar{r}_{i'k}t)\le 0 \qquad \forall t\in\mathcal{T}_k, \forall k\in[K] && \text{(57b)}\\
&\qquad\sup_{t\in\mathcal{T}_k}\sup_{\boldsymbol{\xi}\in\Xi_\xi} -\boldsymbol{a}_n(\boldsymbol{x})^\top\boldsymbol{\xi} - b_n(\boldsymbol{x}) + c_n t - \lambda_k^1\|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_i\| \le q_{ik}+\bar{q}_{ik}t \quad \forall i\in[M_\xi], n\in[N+1], k\in[K] && \text{(57c)}\\
&\qquad\sup_{t\in\mathcal{T}_k}\sup_{\boldsymbol{\zeta}\in\Xi_\zeta}\min_{n\in[N+1]} \boldsymbol{a}_n^{0\top}\boldsymbol{\zeta} + b_n^0 - c_n^0 t - \lambda_k^2\|\boldsymbol{\zeta}-\hat{\boldsymbol{\zeta}}_{i'}\| \le r_{i'k}+\bar{r}_{i'k}t \quad \forall i'\in[M_\zeta], k\in[K] && \text{(57d)}\\
&\qquad\boldsymbol{x}\in\mathcal{X}; \boldsymbol{\lambda}^1,\boldsymbol{\lambda}^2\ge 0; \boldsymbol{q},\bar{\boldsymbol{q}}\in\mathbb{R}^{M_\xi\times K}; \boldsymbol{r},\bar{\boldsymbol{r}}\in\mathbb{R}^{M_\zeta\times K}. && \text{(57e)}
\end{aligned}$$

where $a_{N+1} = b_{N+1} = c_{N+1} = a_{N+1}^0 = b_{N+1}^0 = c_{N+1} = 0$, and $c_n = c_n^0 = 1$ for all $n\in[N]$.

First, as in the proof of Theorem 2, for constraint (57b), one can directly derive the following equivalent formulation which verifies the two boundary scenarios $\bar{t}_k^-$ and $\bar{t}_k^+$ given that the constraint functions are linear in $t$,

$$\lambda_k^1 \epsilon_\xi + \lambda_k^2 \epsilon_\zeta + \frac{1}{M_\xi} \sum_{i \in [M_\xi]} (\bar{q}_{ik} + q_{ik} t_k^+) + \frac{1}{M_\zeta} \sum_{i' \in [M_\zeta]} (\bar{r}_{i'k} + r_{i'k} t_k^+) \leq 0 \qquad \forall k \in [K] \tag{58}$$

$$\lambda_k^1 \epsilon_\xi + \lambda_k^2 \epsilon_\zeta + \frac{1}{M_\xi} \sum_{i \in [M_\xi]} (\bar{q}_{ik} + q_{ik} t_k^-) + \frac{1}{M_\zeta} \sum_{i' \in [M_\zeta]} (\bar{r}_{i'k} + r_{i'k} t_k^-) \leq 0 \qquad \forall k \in [K]. \tag{59}$$

Next, we can treat each of constraints (57c) and (57d) as constraint (10c) that was treated in the proof of Theorem 2 to derive the tractable reformulations of the constraints in problem (57). Letting $g_n(\boldsymbol{x}, \boldsymbol{\xi}, t) := -\boldsymbol{a}_n(\boldsymbol{x})^\top \boldsymbol{\xi} - b_n(\boldsymbol{x}) + c_n t$, we have that constraint (57c) is equivalent to the condition that there exists $\boldsymbol{u} \in \mathbb{R}^{M_\xi \times K \times (N+1)}$, and $\boldsymbol{w} \in \mathbb{R}^{M_\xi \times K \times (N+1) \times m}$ such that:

$$\delta\left(\boldsymbol{w}_{ikn} - \boldsymbol{a}_n(\boldsymbol{x}) \mid \Xi_\xi\right) - {\boldsymbol{w}_{ikn}}^\top \hat{\boldsymbol{\xi}}_i + \bar{t}_k^+ u_{ikn} - \bar{q}_{ik} - b_n(\boldsymbol{x}) \leq 0 \qquad \forall i \in [M_\xi], n \in [N+1], k \in [K] \tag{60}$$

$$\delta\left(\boldsymbol{w}_{ikn} - \boldsymbol{a}_n(\boldsymbol{x}) \mid \Xi_\xi\right) - {\boldsymbol{w}_{ikn}}^\top \hat{\boldsymbol{\xi}}_i + \bar{t}_k^- u_{ikn} - \bar{q}_{ik} - b_n(\boldsymbol{x}) \leq 0 \qquad \forall i \in [M_\xi], n \in [N+1], k \in [K] \tag{61}$$

$$\|\boldsymbol{w}_{ink}\|_* \leq \lambda_k^1 \qquad \forall i \in [M_\xi], n \in [N+1], k \in [K] \tag{62}$$

$$u_{ikn} + q_{ik} - c_n = 0 \qquad \forall i \in [M_\xi], n \in [N+1], k \in [K]. \tag{63}$$

Similarly, for constraint (57d), we can let $g_n(\boldsymbol{x}, \boldsymbol{\zeta}, t) := \inf\limits_{\boldsymbol{\rho} \geq 0: \sum_{n'} \rho_{n'} \leq 1} \sum\limits_{n' \in [N]} \rho_{n'} \left( {\boldsymbol{a}_{n'}^0}^\top \boldsymbol{\xi} + b_{n'}^0 - t \right)$ to obtain that it is equivalent to the condition that there exists $\boldsymbol{u} \in \mathbb{R}^{M_\zeta \times K}$, $\boldsymbol{w} \in \mathbb{R}^{M_\zeta \times K \times m}$, and $\boldsymbol{\rho} \in \mathbb{R}^{M_\zeta \times K \times N}$ such that:

$$\delta(\boldsymbol{w}_{i'k} + \sum_{n' \in [N]} \rho_{ikn'} \boldsymbol{a}_{n'}^0 \mid \Xi_\zeta) - \boldsymbol{w}_{i'k}^\top \hat{\boldsymbol{\zeta}}_{i'} + \bar{t}_k^+ u_{i'k} - \bar{r}_{i'k} + \sum_{n' \in [N]} \rho_{ikn'} b_{n'}^0 \leq 0 \qquad \forall i' \in [M_\zeta], k \in [K] \tag{64}$$

$$\delta(\boldsymbol{w}_{i'k} + \sum_{n' \in [N]} \rho_{ikn'} \boldsymbol{a}_{n'}^0 \mid \Xi_\zeta) - \boldsymbol{w}_{i'k}^\top \hat{\boldsymbol{\zeta}}_{i'} + \bar{t}_k^- u_{i'k} - \bar{r}_{i'k} + \sum_{n' \in [N]} \rho_{ikn'} b_{n'}^0 \leq 0 \qquad \forall i' \in [M_\zeta], k \in [K] \tag{65}$$

$$\|\boldsymbol{w}_{i'k}\|_* \leq \lambda_k^2 \qquad \forall i' \in [M_\zeta], k \in [K] \tag{66}$$

$$u_{i'k} + r_{i'k} + \sum_{n' \in [N]} \rho_{ikn'} = 0 \qquad \forall i' \in [M_\zeta], k \in [K] \tag{67}$$

$$\sum_{n' \in [N]} \rho_{ikn'} \leq 1 \qquad \forall i' \in [M_\zeta], k \in [K] \tag{68}$$

$$\boldsymbol{\rho} \geq 0. \tag{69}$$

Finally, this completes our finite-dimensional convex optimization formulation problem (15), which is further equivalent to a linear programming problem if Assumption 3 holds and when $\mathcal{X}$, $\Xi_\xi$ and $\Xi_\zeta$ are polyhedral.

### B.5. Out-of-Sample Distance from SSD Feasibility

In this appendix, we present a method for measuring the distance from SSD feasibility when the controlled and reference are described using their distribution function. In the context of out-of-sample analysis, the two distribution functions will constitute of the empirical distribution function based on out-of-sample observations considered equiprobable.

We first propose a definition of distance to SSD feasibility for distribution functions based on Wasserstein distance in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We then show how to mathematically calculate this distance when the distribution functions are empirical ones.

Definition 5. Numerically speaking, given two random earning variables $X$ and $Y$, and SSD constraint $X \succeq_{(2)} Y$, the type-$r$ Wasserstein distance of distribution function $F_X$ from the reference $F_Y$ is defined as $\text{dist}_{\text{SSD}}(F_X; F_Y)$, where $F_X$ is the distribution function of $X$ and where

$$\begin{aligned} \text{dist}_{\text{SSD}}(F_X; F_Y) := \quad & \inf_{(\hat{\mathbb{P}}, \bar{\mathbb{P}}, \mathbb{G}) \in \mathcal{M}(\mathbb{R})^3} d_{\text{W}}^r(\hat{\mathbb{P}}, \mathbb{G}) \\ & \text{subject to } \xi \succeq_{(2)}^{(\xi, \bar{\xi}) \sim \mathbb{G} \times \bar{\mathbb{P}}} \bar{\xi} \\ & \qquad \hat{\mathbb{P}}(\hat{\xi} \le z) = F_X(z), \ \forall z \in \mathbb{R} \\ & \qquad \bar{\mathbb{P}}(\bar{\xi} \le z) = F_Y(z), \ \forall z \in \mathbb{R}, \end{aligned}$$

where $\mathcal{M}(\mathbb{R})^3$ is the Cartesian product of three copies of the set of all probability measures in the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and where $d_{\text{W}}^r$ is the type-$r$ Wasserstein metric on $\mathcal{M}(\mathbb{R})$. In other words, $\text{dist}_{\text{SSD}}(F_X; F_Y)$ measures the smallest amount of mass (measured in Wasserstein distance on $\mathbb{R}$) that needs to be moved in order to make the distribution function of $X$ stochastically dominate the distribution function of $Y$.

Given two random variables $\hat{X}$ and $\bar{X}$, we focus on the case where both $F_{\hat{X}}$ and $F_{\bar{X}}$ are empirical distributions based on $M$ samples, i.e., $F_{\hat{X}}(x) := \sum_{i=1}^{M} \mathbf{1}\{\hat{x}_i \le x\}$ and $F_{\bar{X}}(x) := \sum_{i=1}^{M} \mathbf{1}\{\bar{x}_i \le x\}$. In this case:

$$\text{dist}_{\text{SSD}}(F_{\hat{X}}; F_{\bar{X}}) = \underset{\{\mathbb{G}_i\}_{i=1}^M \in \mathcal{M}(\mathbb{R})^M}{\text{minimize}} \left( 1/M \sum_{i \in [M]} \mathbb{E}_{\mathbb{G}_i}[\|\xi - \hat{x}_i\|^r] \right)^{1/r} \tag{70a}$$

$$\text{subject to } 1/M \sum_{i \in [M]} \mathbb{E}_{\mathbb{G}_i}[(t - \xi)^+] \le 1/M \sum_{i \in [M]} (t - \bar{x}_i)^+ \quad \forall t \in \mathbb{R}, \tag{70b}$$

where $\mathbb{E}_{\mathbb{G}_i}[h(\xi)]$ refers to the expected value of $h(\xi)$ when $\xi \sim \mathbb{G}_i$, and where $\mathbb{G}$ is parametrized as $\mathbb{G} = 1/M \sum_i \mathbb{G}_i$ to model with each $\mathbb{G}_i$ how the mass is moved from $\hat{x}_i$ to create the distribution $\mathbb{G}$. By Jensen inequality, we have that for any feasible solution $\{\mathbb{G}_i\}_{i=1}^M$ to problem (70), the candidate

$\{\delta_{\mu_i}\}_{i=1}^M$, where $\delta_x$ refers to the Dirac measure that puts all its mass at $x$ and where $\mu_i := \mathbb{E}_{\mathbb{G}_i}[x_i]$, is also feasible since:

$$1/M \sum_{i\in[M]} \mathbb{E}_{\delta_{\mu_i}}[(t-\xi)^+] = 1/M \sum_{i\in[M]} (t - \mathbb{E}_{\mathbb{G}_i}[\xi])^+ \leq 1/M \sum_{i\in[M]} \mathbb{E}_{\mathbb{G}_i}[(t-\xi)^+] \leq 1/M \sum_{i\in[M]} (t-\bar{x}_i)^+$$

and achieves a lower objective value since again

$$\left(1/M \sum_{i\in[M]} \mathbb{E}_{\delta_{\mu_i}}[\|\xi - \hat{x}_i\|^r]\right)^{1/r} = \left(1/M \sum_{i\in[M]} \|\mathbb{E}_{\mathbb{G}_i}[\xi] - \hat{x}_i\|^r\right)^{1/r} \leq \left(1/M \sum_{i\in[M]} \mathbb{E}_{\mathbb{G}_i}[\|\xi - \hat{x}_i\|^r]\right)^{1/r}.$$

The problem therefore reduces to:

$$\begin{aligned} \mathrm{dist}_{\mathrm{SSD}}(F_{\hat{X}}; F_{\bar{X}}) \;=\; &\underset{\boldsymbol{\mu}}{\text{minimize}} \;\; \left(1/M \sum_{i\in[M]} \|\mu_i - \hat{x}_i\|^r\right)^{1/r} \\ &\text{subject to } 1/M \sum_{i\in[M]} (t-\mu_i)^+ \leq 1/M \sum_{i\in[M]} (t-\bar{x}_i)^+ \qquad \forall t \in \mathbb{R}. \end{aligned}$$

Furthermore, by Proposition 3.2 in Dentcheva and Ruszczynski (2003), it further reduces to the following finite-dimensional convex optimization problem:

$$\mathrm{dist}_{\mathrm{SSD}}(F_{\hat{X}}; F_{\bar{X}}) \;=\; \underset{\boldsymbol{\mu}}{\text{minimize}} \;\; \left(1/M \sum_{i\in[M]} \|\mu_i - \hat{x}_i\|^r\right)^{1/r} \tag{71a}$$

$$\text{subject to } 1/M \sum_{i\in[M]} (\bar{x}_j - \mu_i)^+ \leq 1/M \sum_{i\in[M]} (\bar{x}_j - \bar{x}_i)^+ \qquad \forall j \in [M]. \tag{71b}$$

In the case of $r=1$ and $r=\infty$ this can easily be computed as shown in the following proposition.

Proposition 12. *Let $r=1$, then*

$$dist_{SSD}(F_{\hat{X}}; F_{\bar{X}}) = \left(\max_{j\in[M]} \frac{1}{M}\sum_{i\in[M]} (\bar{x}_j - \hat{x}_i)^+ - \frac{1}{M}\sum_{i\in[M]} (\bar{x}_j - \bar{x}_i)^+\right)^+.$$

*Alternatively, if $r=\infty$, then*

$$dist_{SSD}(F_{\hat{X}}; F_{\bar{X}}) \;=\; \underset{\Delta}{\text{minimize}} \;\; \Delta \tag{72a}$$

$$\text{subject to } 1/M \sum_{i\in[M]} (\bar{x}_j - \Delta - \hat{x}_i)^+ \leq 1/M \sum_{i\in[M]} (\bar{x}_j - \bar{x}_i)^+ \qquad \forall j \in [M] \tag{72b}$$

$$\Delta \in [0, \max_i \bar{x}_i - \min_i \hat{x}_i], \tag{72c}$$

*which can be solved using a bisection on $\Delta$.*

. **Proof of Proposition 12** We start with the case $r=1$. First, we show that

$$\mathrm{dist}_{\mathrm{SSD}}(F_{\hat{X}}; F_{\bar{X}}) \geq \Upsilon := \left(\max_{j\in[M]} \frac{1}{M}\sum_{i\in[M]} (\bar{x}_j - \hat{x}_i)^+ - \frac{1}{M}\sum_{i\in[M]} (\bar{x}_j - \bar{x}_i)^+\right)^+.$$

Namely,

$$
\begin{aligned}
&\begin{cases}
\underset{\boldsymbol{\mu}}{\text{minimize}}\ 1/M \sum_{i\in[M]} |\mu_i - \hat{x}_i| \\
\text{subject to } 1/M \sum_{i\in[M]} (\bar{x}_j - \mu_i)^+ \leq 1/M \sum_{i\in[M]} (\bar{x}_j - \bar{x}_i)^+ \quad \forall j \in [M]
\end{cases} \\
=&\begin{cases}
\underset{\Delta}{\text{minimize}}\ \mathbf{1}/M \sum_{i\in[M]} |\hat{x}_i + \Delta_i - \hat{x}_i| \\
\text{subject to } 1/M \sum_{i\in[M]} (\bar{x}_j - \hat{x}_i - \Delta_i)^+ \leq 1/M \sum_{i\in[M]} (\bar{x}_j - \bar{x}_i)^+ \quad \forall j \in [M]
\end{cases} \\
=&\begin{cases}
\underset{\Delta}{\text{minimize}}\ 1/M \sum_{i\in[M]} |\Delta_i| \\
\text{subject to } 1/M \sum_{i\in[M]} (\bar{x}_j - \hat{x}_i - \Delta_i)^+ \leq 1/M \sum_{i\in[M]} (\bar{x}_j - \bar{x}_i)^+ \quad \forall j \in [M]
\end{cases} \\
\geq&\begin{cases}
\underset{\Delta}{\text{minimize}}\ 1/M \sum_{i\in[M]} |\Delta_i| \\
\text{subject to } 1/M \sum_{i\in[M]} (\bar{x}_j - \hat{x}_i)^+ - |\Delta_i| \leq 1/M \sum_{i\in[M]} (\bar{x}_j - \bar{x}_i)^+ \quad \forall j \in [M]
\end{cases} \\
=&\begin{cases}
\underset{\Delta}{\text{minimize}}\ 1/M \sum_{i\in[M]} |\Delta_i| \\
\text{subject to } 1/M \sum_{i\in[N]} |\Delta_i| \geq 1/M \sum_{i\in[M]} (\bar{x}_j - \hat{x}_i)^+ - 1/M \sum_{i\in[M]} (\bar{x}_j - \bar{x}_i)^+ \quad \forall j \in [M]
\end{cases} \\
=&\Upsilon.
\end{aligned}
$$

We now show that $\text{dist}_{\text{SSD}}(F_{\hat{X}}; F_{\bar{X}}) \leq \Upsilon$ by identifying a solution $\boldsymbol{\mu}$ that is feasible in problem (71) and achieves an objective value of $\Upsilon$. In particular, we set $\mu_i := \max(B, \hat{x}_i)$ for some $B \geq 0$ such that $g(B) := (1/M)\sum_{i\in[M]} |\max(B, \hat{x}_i) - \hat{x}_i| = (1/M)\sum_{i\in[M]} \max(B, \hat{x}_i) - \hat{x}_i = \Upsilon$. This is always possible since $g(B)$ is continuous and non-decreasing, and returns $g(0) = 0$ and $\lim_{B\to\infty} g(B) = \infty$. We are left with showing that constraint (71b) is satisfied. In doing so, we identify to cases. First, the cases where $\bar{x}_j < B$, then the others. If $j$ is such that $\bar{x}_j < B$, then:

$$
1/M \sum_{i\in[M]} (\bar{x}_j - \mu_i)^+ = 1/M \sum_{i\in[M]} (\bar{x}_j - \max(B, \hat{x}_i))^+ \leq 1/M \sum_{i\in[M]} (\bar{x}_j - B)^+ = 0 \leq 1/M \sum_{i\in[M]} (\bar{x}_j - \bar{x}_i)^+.
$$

In the case that $\bar{x}_j > B$,

$$
\begin{aligned}
1/M \sum_{i\in[M]} (\bar{x}_j - \mu_i)^+ &= 1/M \sum_{i\in[M]:\hat{x}_i\geq B} (\bar{x}_j - \hat{x}_i)^+ + 1/M \sum_{i\in[M]:\hat{x}_i<B} (\bar{x}_j - B)^+ \\
&= 1/M \sum_{i\in[M]:\hat{x}_i\geq B} (\bar{x}_j - \hat{x}_i)^+ + 1/M \sum_{i\in[M]:\hat{x}_i<B} (\bar{x}_j - \hat{x}_i + \hat{x}_i - B) \\
&= 1/M \sum_{i\in[M]} (\bar{x}_j - \hat{x}_i)^+ - 1/M \sum_{i\in[M]:\hat{x}_i<B} (B - \hat{x}_i)
\end{aligned}
$$

$$
\begin{aligned}
&= 1/M \sum_{i\in[M]} (\bar{x}_j - \hat{x}_i)^+ - \Upsilon \\
&\leq 1/M \sum_{i\in[M]} (\bar{x}_j - \hat{x}_i)^+ - \frac{1}{M}\sum_{i\in[M]} (\bar{x}_j - \hat{x}_i)^+ + \frac{1}{M}\sum_{i\in[M]} (\bar{x}_j - \bar{x}_i)^+ \\
&= 1/M \sum_{i\in[M]} (\bar{x}_j - \bar{x}_i)^+.
\end{aligned}
$$

Alternatively, in the case of $r=\infty$, problem (71) reduces to (72), which searches for a solution of the form $\mu_i := \hat{x}_i + \Delta$ to problem (71). Such a solution is optimal for (71) since for any solution $\boldsymbol{\mu}$ to problem (71), one can construct the candidate $\boldsymbol{\mu}' := \hat{\boldsymbol{x}} + \|\hat{\boldsymbol{x}} - \boldsymbol{\mu}\|_\infty$ which is also feasible and achieves the same objective value. □