A Data-Driven Approach for Multi-Stage Linear Optimization

Dimitris Bertsimas
Operations Research Center, Massachusetts Institute of Technology, dbertsim@mit.edu

Shimrit Shtern
The William Davidson Faculty of Industrial Engineering & Management, Technion - Israel Institute of Technology, shimrits@technion.ac.il

Bradley Sturt
Operations Research Center, Massachusetts Institute of Technology, bsturt@mit.edu

Multi-stage linear optimization is an integral modeling paradigm in supply chain, energy planning, and finance. However, these problems are computationally challenging, and identifying the correlation structure of the uncertainty across stages presents difficulties. In this paper, we propose a novel data-driven framework for multi-stage linear optimization based on a simple robustification of the data. We show that the framework results in tractable reformulations, even in problems with integer decisions, and provide computational evidence which supports this assertion. Moreover, we establish general convergence guarantees on the average out-of-sample performance of the proposed framework as more data is obtained. To the best of our knowledge, these data-driven guarantees are the first of their kind for multi-stage linear optimization problems with uncertainty that is arbitrarily correlated across stages. We demonstrate the out-of-sample performance and tractability of the proposed approach on data-driven lot sizing and inventory management problems. We also develop theoretical results on a class of Wasserstein ambiguity sets, which are of independent interest.

Key words: data-driven decision making; multi-stage stochastic programming; distributionally robust optimization; adaptive optimization; nonparametric; Wasserstein ambiguity sets.

1. Introduction

In the traditional formulation of linear optimization, one makes a decision which minimizes a known objective function and satisfies a known set of constraints, represented by

\[
\begin{align*}
\text{minimize} & \quad c^T x \\
\text{subject to} & \quad Ax \leq b,
\end{align*}
\]

where \( A \) is a matrix with \( m \) rows and \( n \) columns. Linear optimization has, by all measures, succeeded as a paradigm for modeling and solving numerous real world problems. However, in many practical applications, the objective and constraints are uncertain at the time of decision making. To incorporate uncertainty into the linear optimization framework, Dantzig (1955) proposed partitioning the decision variables into multiple stages, which are made sequentially as more unknown
quantities become observed. This formulation is known today as stochastic multi-stage linear optimization, and has become an integral modeling paradigm in supply chain, energy planning, and finance, among many others; see Birge and Louveaux (2011) and Shapiro et al. (2009).

Unfortunately, stochastic multi-stage linear optimization is known for its computational intractability. Intuitively speaking, the difficulty stems from the need to account for the unfolding of uncertainty and the ability to adapt to it. From a computational complexity standpoint, these problems are \#P-hard, which underscores the challenge of obtaining exact solutions (Dyer and Stougie 2006, Hanasusanto et al. 2016). Thus, much attention has been placed on approximation schemes, including scenario-based decomposition methods (such as Dantzig and Madansky (1961), Birge (1985)), randomized sampling algorithms (Swamy and Shmoys 2012, Shapiro and Nemirovski 2005, Shapiro 2010), scenario reduction methods (Heitsch and Römisch 2009, Pflug and Pichler 2014), and decision rule approximations (Chen et al. 2008, Georghiou et al. 2015, Bodur and Luedtke 2018), among many others.

To tractably address multi-stage linear optimization problems, the alternative paradigm of robust optimization has enjoyed significant success. Originating with Soyster (1973) and Ben-Tal and Nemirovski (1999), the uncertain parameters in robust optimization are chosen adversarially from an uncertainty set, as opposed to having a probability distribution; see Ben-Tal et al. (2009) and Bertsimas et al. (2011). To approximate robust multi-stage linear optimization problems, Ben-Tal et al. (2004) proposed a decision rule approximation, whereby the decisions in each stage are restricted to an affine dependence on the information revealed up to that point, resulting in a highly tractable conic optimization problem. Many practically-tractable decision rule approximation schemes have since been proposed and analyzed for robust multi-stage linear optimization problems (known as adaptive or adjustable optimization), including nonlinear functions via lifting (Chen and Zhang 2009), finite adaptability (Bertsimas and Caramanis 2010, Postek and Hertog 2016, Bertsimas and Dunning 2016), piecewise linear decision rules (Bertsimas and Georghiou 2015, 2018), among many others.

Despite the substantial progress in tractability, a central critique of adaptive optimization is that it does not aim to find solutions which perform well on average. With the goal of combining practical tractability with good average performance, the framework of distributionally robust optimization (DRO) offers significant potential. First proposed by Scarf (1958), DRO models the uncertainty with a probability distribution, but the distribution is presumed unknown and contained in an ambiguity set of distributions. Even though single-stage stochastic optimization is generally intractable, the introduction of ambiguity can surprisingly emit tractable reformulations (Delage and Ye 2010, Goh and Sim 2010, Wiesemann et al. 2014). Consequently, the extension of
DRO to multi-stage decision making is an active area of research; see Xin and Goldberg (2015) for inventory problems with martingale demands and Bertsimas et al. (2018c) for multi-stage linear optimization problems with moment-based ambiguity sets.

The recent years have seen significant developments in methodologies for incorporating historical data into the decision making process. For example, several methods have been proposed to incorporate data into the construction of uncertainty sets in robust optimization (Hong et al. 2017, Tulabandhula and Rudin 2014, Bertsimas et al. 2018a). For multi-stage decision making, a few non-parametric approaches have been proposed, including a phi-divergence approach by Klabjan et al. (2013) for a class of single-item stochastic lot-sizing problems, as well as kernel density estimation-based approaches by Pflug and Pichler (2016) and Hanasusanto and Kuhn (2013) for an abstract class of multi-stage problems (without constraints) and dynamic programming, respectively. In addition, there has been a proliferation of data-driven constructions of ambiguity sets for DRO which emit probabilistic guarantees, including those based on the Wasserstein distance (Esfahani and Kuhn 2018, Gao and Kleywegt 2016), phi-divergences (Ben-Tal et al. 2013, Bayraksan and Love 2015, Van Parys et al. 2017), and statistical hypothesis tests (Bertsimas et al. 2018b). Many of these approaches have since been applied to the particular case of two-stage linear optimization problems, including Jiang and Guan (2018) for phi-divergence and Hanasusanto and Kuhn (2018), Zhao and Guan (2018), Chen et al. (2017) for Wasserstein ambiguity sets. To the best of our knowledge, there are no previous data-driven approaches for multi-stage linear optimization (with more than two stages) that combine practical tractability with nonparametric performance guarantees.

This paper aspires to revisit multi-stage linear optimization in the data-driven era. Our main contribution is a novel data-driven framework for multi-stage linear optimization based on a simple robustification of the data. Leveraging methods from adaptive optimization, we show that the framework results in tractable reformulations, even in problems with integer decisions, and provide computational evidence which supports this assertion. Furthermore, utilizing recent and new results on DRO with a Wasserstein ambiguity set, we show the framework enjoys nonparametric guarantees on the average performance which are, to the best of our knowledge, the first of their kind for stochastic multi-stage linear optimization. Since real-life applications typically start with historical data, the proposed framework and tractable reformulations are applicable to settings regularly found in supply chain, energy planning, and finance, among many others.

The key results of this paper are summarized as follows.

- We propose a general data-driven framework for addressing multi-stage linear optimization problems. Specifically, we construct an uncertainty set around each historical sample path. We
then evaluate decision rules by averaging over the worst-case realization from each uncertainty set. In contrast to adaptive optimization, we average over multiple uncertainty set and thus aim to find solutions with good average performance. For a particular choice of uncertainty sets, we also show that the proposed framework is equivalent to a distributionally robust optimization problem using an $\infty$-Wasserstein ambiguity set, which we show emits properties that are essential for good performance in multi-stage settings.

- We present a practically tractable algorithm for finding high quality decision rules to the proposed data-driven framework. The approach is based on an extension of finite adaptability from adaptive optimization, wherein the uncertainty sets are partitioned into smaller regions, and a separate static or linear decision rule is optimized for each region. Importantly, by exploiting the duality of the proposed data-driven framework, we obtain a compact reformulation for finite adaptability which scales efficiently in the number of data points. The approach readily extends to problems with integer decision variables, and the practitioner can trade off the tightness of their approximations with an increase in computational cost.

- We establish nonparametric convergence guarantees for the proposed data-driven framework. Specifically, when the historical data is drawn independently from a joint probability distribution, we prove that the optimal objective value and feasible decision rules to the proposed data-driven framework converge almost surely to those of the underlying stochastic multi-stage linear optimization problem as the number of data points tends to infinity. Importantly, these convergence guarantees hold for probability distributions with any correlation structure across the stages. To the best of our knowledge, such nonparametric data-driven guarantees are first-of-kind for multi-stage linear optimization, and provide assurance of the proposed framework performance in diverse real-world applications. The proofs leverage new results on DRO with $\infty$-Wasserstein ambiguity sets, which are of independent interest.

- We examine the quality of decision rules and running times of our method in various applications. First, we consider a mixed-integer lot-sizing problem for managing the inventory of a short lifecycle product, where our data consists of historical demands for similar products. Second, we consider a classic multi-stage inventory problem with autoregressive demands. In both cases, the demands are correlated between stages, and the joint distributions are challenging to identify from historical data. Using the proposed methods, we obtain decision rules with high-quality performance (which are nearly optimal) in minutes for problems with up to ten stages and hundreds of data points.

Our paper is organized as follows. Section 2 provides the formulation of multi-stage linear optimization, describes the data-driven setting of interest, and presents the new data-driven framework.
In Section 3, we present the approximation algorithm based on finite adaptability. The relationships between the proposed approach and DRO, along with a discussion of possible alternative approaches, are found in Section 4. The convergence guarantees are described in Section 5. In Section 6, we discuss numerical experiments of the proposed algorithm and comparisons to existing approaches. We conclude in Section 7. All technical proofs are relegated to the attached electronic companion.

**Notation.** We denote the real numbers by $\mathbb{R}$ and the integers by $\mathbb{Z}$. Lowercase bold letters $\mathbf{v} \in \mathbb{R}^d$ denote a vector, and $v_i$ denotes its $i$-th element. We similarly use uppercase bold letters $\mathbf{M} \in \mathbb{R}^{d \times k}$ for matrices. Given vectors $\mathbf{v}, \mathbf{u} \in \mathbb{R}^d$, we let both $\mathbf{v} \cdot \mathbf{u}$ and $\mathbf{v}^\top \mathbf{u}$ denote $v_1u_1 + \cdots + v_du_d$. We let $\mathbf{0}, e, \infty, -\infty$ denote vectors of all zeros, ones, infinity, and minus infinity, respectively. When $K \in \mathbb{N}$, $[K]$ is the set $\{1, \ldots, K\}$; however, when $\mathbf{v} \in \mathbb{R}^d$, we let $[\mathbf{v}]_+$ be the vector $(\max\{v_1,0\}, \ldots, \max\{v_d,0\})$. The operator $\|\cdot\|$ denotes an arbitrary norm in $\mathbb{R}^d$ unless otherwise stated. Given a set $P \subseteq \mathbb{R}^d$, the interior is denoted by $\text{int}(P)$ and the empty set by $\emptyset$. We let $\mathcal{P}(Z)$ denote the space of Borel probability distributions $\mathbb{P}$ with support contained in $Z \subseteq \mathbb{R}^d$, whereby $\mathbb{P}(\xi \in A)$ is the probability that a random vector $\xi$ is contained in the Borel set $A \subseteq Z$ and $\mathbb{E}_\mathbb{P}[f(\xi)] = \int_Z f(\xi) d\mathbb{P}(\xi)$. We let $\mathbb{I}\{\cdot\}$ denote the indicator function, where $\mathbb{I}\{\xi \in A\} = 1$ if $\xi \in A$ and 0 otherwise.

### 2. The New Data-Driven Approach

In this section, we propose the novel data-driven framework for multi-stage linear optimization. In Sections 2.1 and 2.2, we review the formulation of multi-stage linear optimization problems and present the data-driven setting. In Section 2.3, we present the novel data-driven framework.

#### 2.1. Multi-Stage Linear Optimization

We consider multi-stage linear optimization problems with $T \geq 1$ stages. The decisions over the time horizon are denoted by $\mathbf{x} \equiv (\mathbf{x}_1, \ldots, \mathbf{x}_T) \in \mathcal{X}$, where $\mathbf{x}_t$ is an $n_t$-dimensional decision vector and the set $\mathcal{X}$ enforces which components of the decisions are continuous and integral. We represent the uncertain parameters observed over the time horizon by $\xi \equiv (\xi_1, \ldots, \xi_T) \in \mathbb{R}^d$, where $\xi_t$ is the $d_t$-dimensional uncertain quantity that is observed immediately after the decision $\mathbf{x}_t$ is selected. For example, the uncertain quantities in each stage may represent the demand for a product in different weeks. Each decision is made as a function of the information observed up to that point, and this dependence is captured by a decision rule of the form

$$\mathbf{x}(\xi) \equiv (\mathbf{x}_1, \mathbf{x}_2(\xi_1), \ldots, \mathbf{x}_T(\xi_1, \ldots, \xi_{T-1})).$$
In multi-stage linear optimization, the goal is to find a decision rule which minimizes a linear cost function and satisfies a set of linear inequalities. These problems are represented compactly by
\[
\begin{align*}
\text{minimize} \quad & c(\xi)^T x(\xi) \\
\text{subject to} \quad & A(\xi) x(\xi) \leq b(\xi).
\end{align*}
\]

In the stochastic formulation, the uncertain quantities \(\xi_1, \ldots, \xi_T\) are modeled as joint random variables, the objective function is minimized in expectation and the constraints must be satisfied almost surely. We shall make the standard assumption that the problem parameters \(A(\xi) \in \mathbb{R}^{m \times n}, b(\xi) \in \mathbb{R}^m,\) and \(c(\xi) \in \mathbb{R}^n\) are linear functions of the form \(b(\xi) \equiv b^0 + B_\xi, c(\xi) \equiv c^0 + C_\xi,\) and denote the \(i\)th row of \(A(\xi)\) by \(a_i(\xi) \equiv a^0_i + A^i_\xi.\)

### 2.2. The Data-Driven Setting

In real life applications, the only knowledge on the uncertainty comes from historical data, and our goal is to leverage the historical data in order to find decision rules which perform well on future realizations of \(\xi.\) We represent the historical data by
\[
\hat{\xi}^j = (\hat{\xi}_1^j, \ldots, \hat{\xi}_T^j), \quad j = 1, \ldots, N,
\]
and refer to each historical realization \(\hat{\xi}^j\) as a sample path. This setting corresponds to many real-life applications. For example, when managing the inventory of a new short lifecycle product, in which the retailer must manage the inventory level over the product’s lifecycle, each sample path may represent the historical sales data over the lifecycle of a previous similar product (Ban et al. 2018, Hu et al. 2017). Another example is energy planning, where operators must coordinate and commit to production levels throughout a day, the output of wind turbines are subject to uncertain weather conditions, and data on historical daily wind patterns is increasingly available (Lorca et al. 2016, Potter et al. 2008).

In addition, we suppose that there may be knowledge of a set \(\Xi \subseteq \mathbb{R}^d\) which contains the historical data and will contain any conceivable future realization of \(\xi.\) For example, when the uncertainty is the demand for a new product or energy produced by a wind turbine, one can typically assume that the uncertainty will be nonnegative, and can thus set \(\Xi\) to \(\mathbb{R}^d_+.\) If no such support information is known, one can set \(\Xi\) to \(\mathbb{R}^d.\)

We assume throughout that the uncertain parameters \(\xi \equiv (\xi_1, \ldots, \xi_T)\) are exogenous of the decisions. That is, there is no mechanism by which the choice of the decisions will impact the future uncertainty. However, we do not preclude the possibility that the uncertain parameters \(\xi_1, \ldots, \xi_T\) are correlated across stages. For example, the sales of a new product over the first several weeks of its lifecycle may be predictive of its sales in the remaining weeks. We will not impose any assumptions on the correlation structure of the uncertain parameters.
2.3. The New Data-Driven Framework

In this section, we present the novel data-driven framework for multi-stage linear optimization.

The proposed framework is the following. First, we construct an uncertainty set $\mathcal{U}_N^j$ around each historical sample path $\bar{\xi}^j \equiv (\bar{\xi}_1^j, \ldots, \bar{\xi}_T^j)$, consisting of all realizations $\zeta \equiv (\zeta_1, \ldots, \zeta_T)$ which are slight perturbations of the historical trajectory. Then, we evaluate decision rules by averaging over the worst-case realization from each uncertainty set, and require that the decision rule is feasible for all realizations in all of the uncertainty sets. Formally, the proposed approach is the following:

$$\begin{align*}
\min_{x \in \mathcal{X}} & \quad \frac{1}{N} \sum_{j=1}^N \sup_{\zeta \in \mathcal{U}_N^j} c(\zeta)^T x(\zeta) \\
\text{subject to} & \quad A(\zeta)x(\zeta) \leq b(\zeta) \\
& \quad \forall \zeta \in \bigcup_{j=1}^N \mathcal{U}_N^j.
\end{align*}$$

(1)

Unlike adaptive optimization, the proposed framework averages over multiple uncertainty sets. Thus, the explicit goal here is to obtain solutions which perform well on average while simultaneously not overfitting the historical data. Interestingly, we note that Problem (1) does not necessitate that decision rules are feasible for every realization in $\Xi$. The feasibility requirements in Problem (1) are justified when the union of the uncertainty sets encompasses the variability of future realizations of the uncertainty.

Out of the various possible constructions of the uncertainty sets, our investigation is focused on uncertainty sets constructed as balls of the form

$$\mathcal{U}_N^j \equiv \left\{ \zeta \in \Xi : \|\zeta - \bar{\xi}^j\| \leq \epsilon_N \right\},$$

where $\epsilon_N \geq 0$ is the “robustness” parameter, chosen by the practitioner, that controls the size of the uncertainty sets. The norm $\|\zeta - \bar{\xi}^j\|$ represents the distance between the two trajectories $\zeta$ and $\bar{\xi}^j$. The rationale for this particular uncertainty set is three-fold. First, it is conceptually simple, requiring only a single parameter $\epsilon_N$, and does not assume any parametric structure on the uncertainty. Second, the uncertainty sets are of similar structure, which can be exploited to obtain tractable reformulations (see Section 3). Finally, under certain probabilistic assumptions, we show that Problem (1) with these uncertainty sets has desirable nonparametric out-of-sample performance guarantees (see Section 5). Using these uncertainty sets, we remark that Problem (1) can be viewed as a DRO problem using an $\infty$-Wasserstein ambiguity set. We discuss this relationship in Section 4, where we compare Problem (1) to possible alternatives.
3. Tractable Approximations via Finite Adaptability

In this section, we develop an algorithm for finding high-quality decision rules for Problem (1), even when decisions are integral. The proposed approach is an extension of finite adaptability from adaptive optimization, wherein decision rules are restricted to separate static or linear decision rules over a partitioned uncertainty set. By exploiting the dual structure of Problem (1), this approximation scheme can be solved by a compact deterministic optimization problem which scales lightly in the number of data points. Furthermore, the practitioner can trade off the tightness of their approximation by considering more granular partitions. The tractability and approximation quality of the proposed algorithm is exemplified through empirical experiments in Section 6.

In Section 3.1, we extend finite adaptability to Problem (1). In Sections 3.2 and 3.3, we develop tractable reformulations for solving the resulting semi-infinite optimization problems. In Section 3.4, we describe an algorithm for partitioning the uncertainty sets.

3.1. Finite Adaptability for Problem (1)

Intuitively speaking, multi-stage decision problems are challenging due to the optimization over an unrestricted space of decision rules. To overcome this difficulty, a common approximation technique is to restrict the decision rules to a space which can more easily be optimized. One such decision rule approximation is finite adaptability (Bertsimas and Caramanis 2010). In this approach, one partitions the uncertainty set into different regions, and optimizes a separate static decision rule for each region.

More formally, let $P_1, \ldots, P_K \subseteq \mathbb{R}^d$ be given sets. We say that these sets form a partition of $\Xi \subseteq \mathbb{R}^d$ if

$$\bigcup_{k \in [K]} P_k = \Xi,$$

and

$$\text{int}(P_k) \cap \text{int}(P_{k'}) = \emptyset \quad \text{for all } k \neq k' \in [K].$$

In finite adaptability, one restricts the space of decision rules to piecewise static functions of the form

$$x(\zeta) = \begin{cases} x_1, & \text{if } \zeta \in P_1, \\ \vdots \\ x_K, & \text{if } \zeta \in P_K, \end{cases}$$

where $x_k \equiv (x^k_1, \ldots, x^k_T) \in X$ for each $k \in [K]$, and $\zeta \equiv (\zeta_1, \ldots, \zeta_T) \in \Xi$ denotes any realization of the uncertainty.
A complication of finite adaptability is that one may not have enough information at any intermediary stage to determine which set $P_k$ will contain the entire trajectory. In other words, at the start of stage $t$, a decision must be chosen after only observing the values of $\xi_{1:t-1} \equiv (\xi_1, \ldots, \xi_{t-1})$, and there may be two or more sets in the partition which cannot yet be distinguished. Fortunately, this complication is avoided by enforcing that $x_k^t = x_{k'}^t$ whenever the corresponding sets overlap in the first $t-1$ stages, which is formalized as follows.

**Proposition 1 (Proposition 4, Bertsimas and Dunning (2016)).** If there exists $\zeta \equiv (\zeta_1, \ldots, \zeta_T) \in P_k$ and $\zeta' \equiv (\zeta'_1, \ldots, \zeta'_T) \in P_{k'}$ such that $\xi_{1:t-1} = \xi'_{1:t-1}$, and $\zeta \in \text{int}(P_k)$ or $\zeta' \in \text{int}(P_{k'})$ hold, then we must enforce the constraint that $x_k^t = x_{k'}^t$ at time stage $t$ as the two sets cannot be distinguished with the uncertain parameters realized by that time stage. Otherwise, we do not need to enforce any constraints at time stage $t$ for this pair.

Thus, Proposition 1 implies that the aforementioned complication is resolved by adding constraints of the form $x_k^t = x_{k'}^t$ for every $(k,k',t)$ such that $P_k$ and $P_{k'}$ are indistinguishable at stage $t$. For brevity, we denote the collection of these tuples $(k,k',t)$ by $\mathcal{T}(P_1, \ldots, P_K)$, which we assume can be tractably computed.

We now extend the approach of finite adaptability to Problem (1). Let $P_1, \ldots, P_K$ be a given partition of $\Xi$, and for each sample path $\hat{\xi}^j$ let us define

$$K_j \equiv \{k \in [K] : \mathcal{U}_{N} \cap P_k \neq \emptyset\}$$

as the indices of sets $P_k$ which intersect the uncertainty set $\mathcal{U}_N$. Then, applying finite adaptability to the epigraph form of Problem (1), we obtain the following approximation of Problem (1):

$$\begin{align*}
\text{minimize} & \quad x^1, \ldots, x^K, v \in \mathbb{R}^N, \frac{1}{N} \sum_{j=1}^{N} v_j \\
\text{subject to} & \quad c(\zeta) \top x_k \leq v_j \\
& \quad A(\zeta)x_k \leq b(\zeta) \quad \forall \zeta \in \mathcal{U}_N \cap P_k, j \in [N], k \in K_j \\
& \quad x_k^t = x_{k'}^t \quad (k,k',t) \in \mathcal{T}(P_1, \ldots, P_K).
\end{align*}$$

We remark that Problem (2) will always result in an upper bound on Problem (1), as the former restricts the space of decision rules to those which are piecewise static over the sets $P_1, \ldots, P_K$. Speaking intuitively, the approximation gap between Problems (1) and (2) depends on the selection and granularity of the partition. By partitioning $\Xi$ into a greater number of sets, Problem (2) may result in a tighter approximation of Problem (1), although this comes with an increase in problem size. The task of choosing the partition in Problem (2) is discussed in Section 3.4.
As stated thus far, Problem (2) has semi-infinite constraints, which must be eliminated in order for the optimization problem to be solvable by off-the-shelf solvers. Importantly, in order to be solvable in practical running times, the size of an equivalent finite-dimensional optimization problem must scale efficiently in both the number of sample paths $N$ and the number of partition sets $K$. In the following section, we present a simple duality argument that allows Problem (2) to be reformulated as a finite-dimensional linear optimization problem which scales lightly in $N$ and $K$.

### 3.2. Tractable Reformulations Via Duality

In this section, we propose a reformulation of Problem (2) as a finite-dimensional linear optimization problem. Specifically, we demonstrate that, under an appropriate construction of the partition and choice of $\Xi$, we can exploit the structure of Problem (2) to obtain an efficient reformulation. The central idea enabling this reformulation follows from the observation that the worst-case realizations over the various uncertainty sets are found by optimizing over identical linear functions. Thus, when constructing the robust counterparts, we can combine the dual decision variables from different uncertainty sets, resulting in a reformulation where the number of auxiliary variables is independent of $N$. This idea is formalized as follows.

**Proposition 2.** Let $\Xi$ and $P_1, \ldots, P_K$ be hyperrectangles of the form $\{\zeta \in \mathbb{R}^d : \ell \leq \zeta \leq u\}$ for $-\infty \leq \ell \leq u \leq \infty$, and let the uncertainty sets $U^N_1, \ldots, U^N_N$ from Section 2.3 be defined with the $\ell_\infty$ norm. Then, Problem (2) can be reformulated by adding at most $O(Km d)$ additional continuous decision variables and $O(m \sum_{j=1}^N |K_j| + Km d)$ additional linear constraints. The reformulation is

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{N} \sum_{j=1}^N v_j \\
\text{subject to} & \quad u_{jk}^i \cdot \mu_{0k}^i - \ell_{jk}^i \cdot \lambda_{0k}^i \leq v_j - c_0^j \cdot x_k^j \quad j \in [N], \ k \in K_j \\
& \quad u_{jk}^i \cdot \mu_{0k}^i - \ell_{jk}^i \cdot \lambda_{0k}^i \leq \ell_{jk}^i - a_i^0 \cdot x_k^j \quad j \in [N], \ k \in K_j, \ i \in [m] \\
& \quad \mu_{0k}^i - \lambda_{0k}^i = C^\top x_k^i \quad k \in [K] \\
& \quad \mu_{0k}^i - \lambda_{0k}^i = (A_i^0)^\top x_k^i - b_i^j \quad k \in [K], \ i \in [m] \\
& \quad x_k^j = x_{k'}^j \quad (k, k', t) \in \mathcal{T}(P_1, \ldots, P_K),
\end{align*}
\]

where $b_i^j$ is the $i$th row of matrix $B$, and $\ell_{ik}, u_{ik}$ are fixed vectors that satisfy

\[
\{\xi \in \mathbb{R}^d : \ell_{ik} \leq \xi \leq u_{ik}\} = U^N_i \cap P_k.
\]

We remark that this reformulation holds even when $\mathcal{X}$ enforces that some decisions are integral.
We now reformulate each of these semi-infinite constraints by introducing auxiliary variables. For Problem (2), we can rewrite it as

\[ \text{Problem (2) can be rewritten as} \]

\[ (3.3. \text{Improved Approximations when } A(\zeta) \text{ and } c(\zeta) \text{ are Constant}) \]

We now consider a specific case of Problem (2) where \( A(\zeta) \equiv A \) and \( c(\zeta) \equiv c \) are fixed, which encompasses numerous applications (see Section 6). For this class of problems, we show in this section that one can obtain even tighter reformulations and better approximations.

3.3. Improved Approximations when \( A(\zeta) \) and \( c(\zeta) \) are Constant

We first consider finite adaptability with static decision rules, as in Section 3.1. In the present setting, the uncertainty only appears in \( b(\zeta) \). Importantly, the worst-case realizations in the constraints are completely independent of the decision variables \( x^k \). Thus, the worst case realization are those for which \( b(\zeta) \) is minimized along each row. Thus, the problem can be reformulated as

\[ \text{We readily observe that, for each } i \in \{0,1,\ldots,m \}, j \in [N], \text{ and } k \in K_j, \text{ it follows from strong duality for linear optimization that} \]

\[ \begin{align*}
\text{minimize} \quad & u^i_k \cdot \mu - \ell^j_k \cdot \lambda \\
\text{subject to} \quad & \tilde{a}_i^k \cdot \zeta \leq \tilde{b}_i^j \\
& \mu, \lambda \in \mathbb{R}^d_+.
\end{align*} \]

We readily observe that, for each \( i \in \{0,1,\ldots,m \} \) and \( k \in [K] \), the solutions \( \mu = [\tilde{a}_i^k]^+ \) and \( \lambda = [-\tilde{a}_i^k]^+ \) are optimal for the above optimization problem since \( \ell^j_k \leq u^j_k \). Importantly, the index \( j \) is absent from these expressions. Thus, the constraints in Line (4) are satisfied if and only if there exists \( \mu_i^k \in \mathbb{R}^d_+ \) and \( \lambda_i^k \in \mathbb{R}^d_+ \) for each \( i \in \{0,1,\ldots,m \} \) and \( k \in [K] \) which satisfy

\[ \begin{align*}
& u^j_k \cdot \mu_i^k - \ell^j_k \cdot \lambda_i^k \leq \tilde{b}_i^j \quad \forall j \in [N] \text{ such that } k \in K_j \\
& \mu_i^k - \lambda_i^k = \tilde{a}_i^k
\end{align*} \]

Substituting the definitions of \( \tilde{a}_i^k \) and \( \tilde{b}_i^j \), we obtain the desired reformulation. \( \square \)

This result suggests that Problem (1) with finite adaptability is highly tractable, in the sense that the size of the resulting reformulation scales lightly in \( N \). Under the same assumptions of Proposition 1, we remark that \( \ell^j_k, u^j_k, \) and \( T(P_1,\ldots,P_K) \) can be computed efficiently by iterating through each \( j \in [N] \) and \( k \in K_j \).
\[
\begin{align*}
\text{minimize} & \quad \frac{1}{N} \sum_{j=1}^{N} v_j \\
\text{subject to} & \quad c^\top x^k \leq v_j \quad j \in [N], \quad k \in \mathcal{K}_j \\
& \quad A x^k \leq b^k \quad k \in [K] \\
& \quad x^k_t = x^k_{t'} \quad (k, k', t) \in T(P_1, \ldots, P_K).
\end{align*}
\]

where \( \bar{b}^k = \min_{j \in [N]: k \in \mathcal{K}_j} \min_{\zeta \in U^j_N \cap P_k} b_j(\zeta) \) can be precomputed.

Second, we consider tighter approximations in which we allow some decisions to be approximated by linear decision rules. Indeed, in problems where some of the decision variables are continuous, static decision rules are unnecessarily restrictive. Instead, for the continuous components we can restrict the corresponding decision rules to piecewise linear functions, as illustrated by the following optimization problem:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{N} \sum_{j=1}^{N} v_j \\
\text{subject to} & \quad c^\top (x^k + F^k \zeta) \leq v_j \\
& \quad A (x^k + F^k \zeta) \leq b(\zeta) \quad \forall \zeta \in U^j_N \cap P_k, \quad j \in [N], \quad k \in \mathcal{K}_j \\
& \quad x^k_t = x^k_{t'}, \quad F^k_t = F^k_{t'} \quad (k, k', t) \in T(P_1, \ldots, P_K).
\end{align*}
\]

In order to ensure that the decision rules are non-anticipative, we restrict the matrices \( F^k \equiv (F^k_1, \ldots, F^k_T) \) such that the columns of \( F^k_t \) corresponding to \( \zeta, \ldots, \zeta_T \) are set to zero. In contrast to static decision rules, such piecewise linear decision rules enable better approximations. Furthermore, one can obtain a reformulation of Problem (5) with a similar tractability of Proposition 2. Indeed, a tractable reformulation is made possible through the fact that Problem (5) can be transformed into a particular case of Problem (2) with decision variables \( \tilde{x}^k = (x^k, \text{vec}(F^k)) \), \( \tilde{X} = \mathcal{X} \times \mathcal{M} \), \( \tilde{c}(\zeta) = (c, \text{vec}(c^\zeta')) \), and \( \tilde{a}_i(\zeta) = (a_i, \text{vec}(a_i^\zeta')) \) where \( a_i \) is the \( i \)th row of matrix \( A \).

### 3.4. An Iterative Approach for Selecting Partitions

Up to this point, we have not addressed how to choose the partition \( P_1, \ldots, P_K \). The selection of these sets is of importance, as they will directly impact the approximation quality and tractability of the resulting optimization problem.

To select partitions, we propose an iterative approach based on schemes presented by Postek and Hertog (2016) and Bertsimas and Dunning (2016). At each iteration, we start with a partition \( P_1, \ldots, P_K \) and solve the resulting approximation of Problem (1) with finite adaptability. We then find the realizations \( \zeta \) in each \( U^j_N \cap P_k \) which are active at the optimal decision rule, and split the
previous partition $P_1, \ldots, P_K$ into more granular sets by separating these worst-case realizations. We present the details and theoretical justification of this heuristic in Section EC.7. The tractability and approximation quality of the proposed iterative approach is demonstrated via empirical experiments in Section 6.

4. Relationships to Distributionally Robust Optimization

In this section, we establish an equivalence of the proposed data-driven framework from Section 2.3 with DRO. Specifically, we show that Problem (1) is equivalent to DRO using the $\infty$-Wasserstein ambiguity set, which has previously received little investigation in the DRO literature. We compare this approach to possible alternatives using the $p$-Wasserstein ambiguity set with $p \in [1, \infty)$, which has received recent consideration and analysis in two-stage settings (see Section 1). In contrast to alternative ambiguity sets, we shall demonstrate that the $\infty$-Wasserstein ambiguity set has a unique property of considering only distributions whose support does not extend too far beyond the historical data, a property which we show is vital for obtaining good out-of-sample performance.

In Section 4.1, we review DRO using Wasserstein ambiguity sets. In Section 4.2, we present the perspective of Problem (1) as a DRO problem using the $\infty$-Wasserstein ambiguity set. In Section 4.3, we propose and analyze an alternative approach using the 1-Wasserstein ambiguity set.

4.1. Background: Wasserstein Ambiguity Sets

The primitive to DRO is an ambiguity set of probability distributions. In data-driven settings, one way to define such an ambiguity set is to consider all the distributions which are not too far from a given nominal distribution. Let $\mathcal{P}(\Xi)$ be the space of probability distributions with support contained in $\Xi$, and let $\hat{P}_N \equiv \frac{1}{N} \sum_{j=1}^{N} \delta_{\xi_j}$ be the empirical distribution of the historical data, which assigns equal weights to all sample paths. Given a metric $d: \mathcal{P}(\Xi) \times \mathcal{P}(\Xi) \to \mathbb{R}^+ \cup \{\infty\}$ between probability distributions, one can define an ambiguity set of the form

$$A = \left\{ Q \in \mathcal{P}(\Xi) : d(Q, \hat{P}_N) \leq \epsilon_N \right\},$$

where $\epsilon_N \geq 0$ is a robustness parameter. Intuitively speaking, this ambiguity set contains all distributions which are close to the empirical distribution. For example, when $\epsilon_N = 0$, such an ambiguity set resolves to contain only the empirical distribution; for larger values of $\epsilon_N$, the ambiguity set includes probability distributions which are further from the empirical distribution.

By choosing different metrics between probability distributions, one obtains different ambiguity sets. In particular, different metrics induce different probabilistic guarantees on the ambiguity set.
Recently, significant attention has been placed on the $p$-Wasserstein distance for $p \in [1, \infty)$, defined as

$$d_p(Q, Q') = \inf \left\{ \left( \int_{\Xi \times \Xi} \|\xi - \xi'\|^p d\Pi(\xi, \xi') \right)^{\frac{1}{p}} : \Pi \text{ is a joint distribution of } \xi \text{ and } \xi' \right\} \text{ with marginals } Q \text{ and } Q', \text{ respectively}.$$

For more details, see, for example, Gao and Kleywegt (2016). Among other reasons, attention has been placed on ambiguity sets using the $p$-Wasserstein distance for $p \in [1, \infty)$ due to its nonparametric probabilistic guarantees; see Esfahani and Kuhn (2018) and Fournier and Guillin (2015).

A particular case of the Wasserstein distance, which is not covered by the aforementioned discussion, is the case where $p = \infty$. Given two probability distributions $Q, Q' \in \mathcal{P}(\Xi)$, their $\infty$-Wasserstein distance is defined by

$$d_\infty(Q, Q') \equiv \inf \left\{ \Pi \text{-ess sup}_{\Xi \times \Xi} \|\xi - \xi'\| : \Pi \text{ is a joint distribution of } \xi \text{ and } \xi' \right\},$$

where the essential supremum of the joint distribution is defined as

$$\Pi \text{-ess sup}_{\Xi \times \Xi} \|\xi - \xi'\| = \inf \left\{ M : \Pi(\|\xi - \xi'\| > M) = 0 \right\}.$$

For more technical details on this distance, we refer the interested reader to Trillos and Šlepčev (2014) and Givens et al. (1984). We henceforth refer to the ambiguity set using a $p$-Wasserstein distance ($p \in [1, \infty]$) as the $p$-Wasserstein ambiguity set.

In contrast to the $p$-Wasserstein ambiguity set for $p \in [1, \infty)$, there are relatively few previous nonparametric probabilistic guarantees on the $\infty$-Wasserstein ambiguity set (see Section 5.1). Nonetheless, we remark that DRO using the $\infty$-Wasserstein ambiguity set has recently received interest with regard to robustness and regularization in machine learning (Gao et al. 2017).

### 4.2. Problem (1) versus DRO with $\infty$-Wasserstein

In order to extend DRO to the multi-stage linear optimization, one can consider problems of the form

$$\begin{align*}
\text{minimize} & \quad x \in X \\
& \sup_{Q \in \mathcal{A}} \mathbb{E}_Q [c(\xi)^\top x(\xi)] \\
\text{subject to} & \quad Q(A(\xi)x(\xi) \leq b(\xi)) = 1 \\
& \quad \forall Q \in \mathcal{A},
\end{align*}$$

(6)

Stated equivalently, this formulation looks for the decision rules which minimize the expected cost and are feasible almost surely with respect to an adversarially chosen probability distribution from the ambiguity set.$^1$

$^1$ One can show that this formulation is equivalent to recent DRO approaches for two-stage linear optimization (Hanasusanto and Kuhn 2018, Chen et al. 2017) and multi-stage linear optimization (Bertsimas et al. 2018c).
We now discuss the proposed data-driven framework of this paper with the Wasserstein ambiguity sets from Section 4.1. The following result shows that Problem (1) can be viewed as a particular case of Problem (6) using the $\infty$-Wasserstein ambiguity set. We begin by presenting an intuitive intermediary result.

**Lemma 1.** The $\infty$-Wasserstein ambiguity set is equivalent to

$$\left\{ \frac{1}{N} \sum_{j=1}^{N} Q_j : Q_j \left( \|\xi - \hat{\xi}^j\| \leq \epsilon_N \right) = 1 \text{ for each } j \in [N] \right\}.$$  

**Proof.** See Section EC.1. □

Therefore, when $\mathcal{A}$ is the $\infty$-Wasserstein ambiguity set, we readily obtain the following result.

**Theorem 1.** Problem (1) is equivalent to Problem (6) using the $\infty$-Wasserstein ambiguity set.

**Proof.** See Section EC.1. □

In order to establish convergence guarantees for Problem (1), we develop several new results on the $\infty$-Wasserstein ambiguity set, which hold under mild probabilistic assumptions (see Section 5).

### 4.3. On the Conservatism of Alternative Wasserstein-based DRO Approaches

We conclude this section by considering an alternative data-driven approach for multi-stage linear optimization: namely, that of Problem (6) using the $p$-Wasserstein ambiguity set for $p \in [1, \infty)$. To the best of our knowledge, such an approach has not previously been proposed for multi-stage linear optimization with more than two stages.

First, we find that Problem (6) with the 1-Wasserstein ambiguity set can be approximated using finite adaptability, similarly to that of Problem (1) in Section 3. We provide a detailed reformulation of this alternative approach in Section EC.8, and discuss the performance of this approach via empirical experiments in Section 6.

However, even if Problem (6) with the 1-Wasserstein ambiguity set could be solved exactly, the constraints induced by this approach are generally overly conservative, as illustrated by the following result.

**Theorem 2.** Let $\mathcal{A}$ be the $p$-Wasserstein ambiguity set with robustness parameter $\epsilon_N > 0$ and $p \in [1, \infty)$. Then, the constraints from Problem (6), i.e.,

$$Q \left( (\mathcal{A}(\xi)x(\xi) \leq b(\xi)) \right) = 1, \quad \forall Q \in \mathcal{A},$$

are satisfied if and only if

$$\mathcal{A}(\zeta)x(\zeta) \leq b(\zeta) \quad \forall \zeta \in \Xi.$$
Proof. See Section EC.2. □

Intuitively speaking, Theorem 2 shows that the \( p \)-Wasserstein ambiguity set (\( p \in [1, \infty) \)) restricts the space of decision rules to those which are feasible for every realization in \( \Xi \). Such a result is in stark contrast to the \( \infty \)-Wasserstein ambiguity set, which only requires feasibility for realizations which are in close proximity to the historical sample paths. The central issue is that \( \Xi \) is not necessarily a tight approximation of the true support and in many cases may be strictly and significantly larger. To illustrate the implications of this conservatism, we present the following example.

**Example 1.** Consider a two-stage problem of the form

\[
\begin{align*}
\text{minimize} & \quad x_1 + \mathbb{E}_x \left[ 4(x_2(\xi_1)) \right] \\
\text{subject to} & \quad x_1 + x_2(\xi_1) \geq \xi_1 \\
& \quad x_2(\xi_1) \leq 1 \\
& \quad x_1, x_2(\xi_1) \geq 0,
\end{align*}
\]

where the constraints must hold almost surely. Let the random variable \( \xi_1 \) be uniformly distributed over \([0, 1]\), in which case the optimal second stage decision rule is \( x_2(\xi_1) = \max\{\xi_1 - x_1, 0\} \), the unique optimal first-stage decision is \( x_1 = \frac{3}{4} \), and the optimal objective value is \( \frac{7}{8} \).

Now suppose that the true distribution of \( \xi_1 \) is unknown, and our only information comes from historical data \( \hat{\xi}_1^1, \ldots, \hat{\xi}_1^N \) and knowledge that the support of \( \xi_1 \) is contained in \( \Xi = [0, \alpha] \) for some specified \( \alpha \geq 1 \). By Theorem (2), solving Problem (6) with the \( p \)-Wasserstein ambiguity set (\( p \in [1, \infty) \)) implies that all constraints must be satisfied for every realization in \( \Xi \). Thus, any feasible solution to Problem (7) must satisfy

\[
\begin{align*}
x_1 + x_2(\alpha) & \geq \alpha \\
x_2(\alpha) & \leq 1.
\end{align*}
\]

These inequalities imply that the space of feasible first-stage decisions is restricted to those which satisfy \( x_1 \geq \alpha - 1 \). If \( \alpha > 7/4 \), then the optimal first-stage decision \( x_1 = 3/4 \) will no longer feasible. If \( \Xi = [0, \infty) \) (by taking the limit as \( \alpha \) goes to infinity), there will be no feasible first-stage decisions. Thus, unless the true support is known with high accuracy by practitioner, the solutions for Problem (7) obtained by solving Problem (6) using the \( p \)-Wasserstein ambiguity set (\( p \in [1, \infty) \)) can be arbitrarily suboptimal, even as more data is obtained. □
The above example demonstrates that any ambiguity set which does not attempt to simultaneously estimate the support of an underlying distribution may result in arbitrarily suboptimal decisions. In contrast to alternatives, the ∞-Wasserstein ambiguity set only contains distributions with support in the vicinity of the historical data. In the next section, we show that this feature of the ∞-Wasserstein ambiguity set facilitates general convergence guarantees for Problem (1).

5. Convergence Guarantees

In this section, we analyze the convergence properties of Problem (1). Specifically, we investigate the behavior of the optimal objective value and feasible decision rules of Problem (1) as the number of historical sample paths tends to infinity. To derive these results, we present new guarantees on the ∞-Wasserstein ambiguity set, which are of independent interest.

In Section 5.1, we introduce necessary notation, discuss related results from the literature, and state our assumptions. In Section 5.2, we present nonparametric convergence guarantees for the feasibility and average performance of decision rules to Problem (1). In Section 5.3, we provide a tight characterization of the optimal objective value of Problem (1).

5.1. Preliminaries and Related Results

Throughout this section, we assume that ξ and ˆξ1, ..., ˆξN are random variables which are independently and identically distributed from a joint distribution P ∈ P(Ξ), where P(Ξ) is the set of probability distributions with P(ξ ∈ Ξ) = 1. Let PN ≡ P × ... × P denote the N-fold distribution on the sample (ˆξ1, ..., ˆξN), and let S ⊆ Ξ denote the support of P; that is, the smallest closed set where P(ξ ∈ S) = 1. We say that a sequence converges “P∞-almost surely” if it converges with probability one over the historical data sequence (ˆξj : j ∈ N).

To the best of our knowledge, relatively few convergence guarantees are known for Problem (1), or for DRO using the ∞-Wasserstein ambiguity set. For instance, the concentration inequalities of Fournier and Guillin (2015), which are used by Esfahani and Kuhn (2018) to establish probabilistic guarantees for the p-Wasserstein ambiguity set when p ∈ [1, ∞), do not readily extend to the case where p = ∞. The closest equivalent nonparametric concentration inequalities for the ∞-Wasserstein distance are by Trillos and Slepčev (2014), which hold only when the underlying distribution is bounded. Other noteworthy results include the convergence guarantees of Xu et al. (2012), which show consistency of machine learning models by averaging over multiple uncertainty sets. Their results require that the objective function is continuous, Ξ ≡ Rd, and the underlying distribution is continuous. Alternatively, Erdoğan and Iyengar (2006) provide feasibility guarantees on robust constraints of a form similar to Problem (1) from the perspective of approximating...
ambiguous chance constraints using the Prohorov metric. Their results can be applied to fixed solutions (Erdo˘gan and Iyengar 2006, Theorems 3,4) or require the constraint functions to have a finite VC-dimension (Erdo˘gan and Iyengar 2006, Theorem 5).

In this paper, our only assumption on the underlying joint distribution is the following.

**Assumption 1.** There exists a constant \( a > 1 \) such that the stochastic process \( \xi \equiv (\xi_1, \ldots, \xi_T) \in \mathbb{R}^d \) satisfies

\[
\mathbb{E}_P \left[ \exp(\|\xi\|^{a}) \right] \equiv \int_{\mathbb{R}^d} \exp(\|\xi\|^{a}) d\mathbb{P}(\xi) < \infty.
\]

As argued by Esfahani and Kuhn (2018), requiring such a light tail on an underlying probability distribution is a mild assumption. Importantly, Assumption 1 does not specify any particular correlation structure on the random variables \( \xi_1, \ldots, \xi_T \) observed over the stages (e.g., independence, Markovian, etc).

Additionally, we will assume that the robustness parameter \( \epsilon_N \geq 0 \) satisfies the following rate.

**Assumption 2.** There exists a constant \( \kappa > 0 \) such that the robustness parameter \( \epsilon_N \geq 0 \) of the uncertainty sets is chosen to be the following:

\[
\epsilon_N \equiv \begin{cases} 
\kappa N^{-\frac{1}{3}}, & \text{if } d = 1, \\
\kappa N^{-\frac{1}{d+1}}, & \text{if } d \geq 2.
\end{cases}
\]

We note that, for many of the following results, the rate on \( \epsilon_N \) from Assumption 2 can be improved; nonetheless, this assumption is made for all of the following results to simplify the exposition.

### 5.2. Convergence Guarantees for Decision Rules

We begin by discussing convergence guarantees for decision rules, which serve two roles. First, the guarantees in this section apply to any decision rule for Problem (1), and thus provide insight for decision rules produced by an approximation algorithm (see Section 3). Second, these results lay the foundation for the convergence guarantees on the optimal objective value of Problem (1) in the following section.

We begin by discussing the out-of-sample feasibility of decision rules for Problem (1). Recall that Problem (1) finds decision rules which are feasible for each realization in the uncertainty sets. However, one cannot guarantee that the decision rule will be feasible for realizations outside of the uncertainty sets. Thus, a pertinent question is whether a decision rule obtained from approximately solving Problem (1) is feasible with high probability.
To address the question of feasibility, we propose leveraging classic results from nonparametric statistics. Let $S_N \equiv \bigcup_{j=1}^N \mathcal{U}_N^j$ be shorthand for the union of the uncertainty sets. We say that a decision rule is $S_N$-feasible if

$$A(\zeta)x(\zeta) \leq b(\zeta), \quad \forall \zeta \in S_N.$$  

In other words, the set of feasible decision rules to Problem (1) are exactly those which are $S_N$-feasible. Our subsequent analysis of utilizes the following (seemingly tautological) observation: for any data set $\hat{\xi}^1, \ldots, \hat{\xi}^N$ and any decision rule that is $S_N$-feasible, we have that

$$P(A(\xi)x(\xi) \leq b(\xi)) \geq P(\xi \in S_N),$$

where $P(\xi \in S_N)$ is shorthand for $P(\xi \in S_N | \hat{\xi}^1, \ldots, \hat{\xi}^N)$. Indeed, this inequality follows from the fact that a decision rule which is $S_N$-feasible is definitionally feasible for all realizations $\zeta \in S_N$, and thus the probability of feasibility is at least the probability that $\xi \in S_N$.

We have thus transformed the analysis of feasible decision rules for Problem (1) to the problem of analyzing the performance of $S_N$ as an estimate of the support $S$. Interestingly, using $S_N$ to estimate the support $S$ of the distribution has been widely studied in the statistics literature, with perhaps the earliest results coming from Devroye and Wise (1980) in detection theory. Since then, the performance of $S_N$ as a non-parametric estimate of $S$ has been studied with applications in cluster analysis and image recognition (Schölkopf et al. 2001, Korostelev and Tsybakov 1993). Using this relationship between Problem (1) and nonparametric support estimation, we obtain the following guarantee on feasibility.

**Theorem 3.** Suppose Assumptions 1 and 2 hold. Then

$$\left(\frac{N^{\frac{1}{d+1}}}{(\log N)^{d+1}}\right)P(\xi \notin S_N) \to 0$$

as $N \to \infty$, $\mathbb{P}^\infty$-almost surely.

**Proof.** See Section EC.3. $\square$

Intuitively speaking, Theorem 3 provides a guarantee that any feasible decision rule to Problem (1) will be feasible with high probability on future data. To illustrate why robustness is indeed necessary to achieve such feasibility guarantees, we consider a simple example from supply chain.

**Example 2.** Consider a manufacturer who must satisfy orders placed by retailers for a product. The orders are received in two phases, $\xi_1$ and $\xi_2$. After observing $\xi_1$, the manufacturer selects a production level at unit cost. After observing $\xi_2$, the manufacturer produces any remaining units
at a per-unit cost of $c > 1$. The goal is to minimize the expected costs, resulting in a three-stage linear optimization problem of the form

$$\begin{align*}
\text{minimize} & \quad \mathbb{E}_P \left[ x_2(\xi_1) + cx_3(\xi_1, \xi_2) \right] \\
\text{subject to} & \quad x_2(\xi_1) + x_3(\xi_1, \xi_2) \geq \xi_1 + \xi_2 \\
& \quad x_2(\xi_1), x_3(\xi_1, \xi_2) \geq 0,
\end{align*}$$

where the constraints must hold almost surely. Let us suppose that $\xi_1$ is a continuous random variable, $\xi_1$ and $\xi_2$ are correlated, and our information consists of historical data $\hat{\xi}_1, \ldots, \hat{\xi}_N$ which are independent and identically distributed random variables with the same distribution as $\xi = (\xi_1, \xi_2)$. Let $\epsilon_N = 0$, in which case Problem (1) takes the form

$$\begin{align*}
\text{minimize} & \quad \frac{1}{N} \sum_{j=1}^N \left( x_2(\hat{\xi}_1^j) + cx_3(\hat{\xi}_1^j, \hat{\xi}_2^j) \right) \\
\text{subject to} & \quad x_2(\hat{\xi}_1^j) + x_3(\hat{\xi}_1^j, \hat{\xi}_2^j) \geq \hat{\xi}_1^j + \hat{\xi}_2^j \\
& \quad x_2(\hat{\xi}_1^j), x_3(\hat{\xi}_1^j, \hat{\xi}_2^j) \geq 0 \quad \forall j \in [N].
\end{align*}$$

Since $\xi_1$ has a continuous distribution, we have that $\hat{\xi}_1^1 \neq \cdots \neq \hat{\xi}_1^N$ almost surely. Thus, an optimal solution for the above optimization problem is

$$x_2(\xi_1) = \begin{cases} 
\hat{\xi}_1^j + \hat{\xi}_2^j, & \text{if } \xi_1 = \hat{\xi}_1^j \text{ for } j \in [N], \\
0, & \text{otherwise;}
\end{cases} \quad x_3(\xi_1, \xi_2) = 0.$$  

However, we readily observe that these decision rules will be infeasible almost surely on out-of-sample data. Thus, a judicious choice of the robustness parameter is necessary to obtain decision rules with good out-of-sample feasibility. □

Next, we establish guarantees on the expected out-of-sample cost of decision rules. In order to prove the convergence guarantees of the optimal objective value in the following section, we desire a general bound on the in-sample cost $\frac{1}{N} \sum_{j=1}^N \sup_{\zeta \in U_N^j} c(\zeta)^\top x(\zeta)$ of a decision rule to its out-of-sample cost $\mathbb{E}_P[ c(\zeta)^\top x(\zeta) ]$. However, the in-sample cost does not capture the behavior of the objective function $c(\zeta)^\top x(\zeta)$ for realizations outside of the uncertainty sets. Thus, we cannot hope to establish a general gap between the in-sample and out-of-sample costs. Therefore, we establish a uniform guarantee which relates the in-sample cost to the out-of-sample cost for realizations which lie in the uncertainty sets.

**Theorem 4.** Suppose Assumptions 1 and 2 hold. Then there exists an $\bar{N} \in \mathbb{N}$, $\mathbb{P}^\infty$-almost surely, such that

$$\mathbb{E}_P \left[ c(\zeta)^\top x(\zeta) \mathbb{I} \{ \zeta \in S_N \} \right] \leq \frac{1}{N} \sum_{j=1}^N \sup_{\zeta \in U_N^j} c(\zeta)^\top x(\zeta) + M_N \sup_{\zeta \in S_N} |c(\zeta)^\top x(\zeta)|$$
for all \( N \geq \bar{N} \) and all decision rules, where 
\[
M_N \equiv N^{-\frac{1}{(d+1)(d+2)}} \log N \to 0 \quad \text{as} \quad N \to \infty.
\]

\textbf{Proof.} See Section EC.4. \( \square \)

The central value of this theorem is that it holds uniformly over all decision rules. Thus, as long as 
\[
|c(\zeta)^\top x(\zeta)|
\]
does not grow too quickly in \( \|\zeta\| \) over some space of decision rules, then the right-most additive term in Theorem 4 will be small.

5.3. Convergence of the Optimal Objective Value

We now shift focus to analyzing the optimal cost of Problem (1). Specifically, we shall show that the optimal objective value of Problem (1) nearly converges \textit{almost surely} to the optimal objective value of a nominal stochastic multi-stage linear optimization problem, providing assurance that Problem (1) will result in near-optimal average performance in big data settings.

In order to prove convergence, we will restrict the space of decision rules to those that satisfy a certain growth condition. Let \( L \geq 0 \) be any fixed constant. Then, we say that a decision rule \( x \in X \) satisfies the growth condition (henceforth referred to as Condition (EXP)) if
\[
|c(\zeta)^\top x(\zeta)| \leq L \exp(\|\zeta\|) \quad \text{for all} \quad \zeta \in \Xi.
\]

(Condition (EXP))

In other words, we say that a decision rule satisfies Condition (EXP) if the cost function resulting from the solution does not grow too quickly for realizations far away from the origin. For shorthand, we let \( x \in X : x-(\text{EXP}) \) be the space of all decision rules which satisfy Condition (EXP).

The results of this section provide bounds on the optimal objective value of Problem (1) with such a restriction on the decision rules. We define the optimal objective value of that optimization problem by \( \hat{J}_N \), where
\[
\hat{J}_N \equiv \minimize_{x \in X : x-(\text{EXP})} \frac{1}{N} \sum_{j=1}^{N} \sup_{\zeta \in \mathcal{U}_j^N} c(\zeta)^\top x(\zeta)
\]
subject to 
\[
A(\zeta)x(\zeta) \leq b(\zeta)
\]
\( \forall \zeta \in \bigcup_{j=1}^{N} \mathcal{U}_j^N. \)

We develop tight lower and upper bounds \( \hat{J}_N \) as the number of sample paths \( N \) tends to infinity.

First, let \( J \) be defined as the maximal objective value of a chance-constrained version of the stochastic multi-stage linear optimization problem.
\[
J \equiv \lim_{\rho \downarrow 0} \minimize_{x \in X : x-(\text{EXP})} \mathbb{E}_{\mathcal{P}} \left[ c(\xi)^\top x(\xi) \right]
\]
subject to 
\[
\mathbb{P} \left( A(\zeta)x(\zeta) \leq b(\zeta) \right) \geq 1 - \rho.
\]
Note that the above limit must exist, as the optimal objective value of the above chance constrained optimization problem is monotone in \(\rho\). Second, let \(\bar{J}\) be the optimal cost of the stochastic multi-stage linear optimization problem with an additional restriction that the decision rules are feasible on an expanded support. Specifically,

\[
\bar{J} \equiv \lim_{\rho \downarrow 0} \min_{\mathbf{x}(\exp) \in \mathcal{X}} \mathbb{E}_\mathcal{P}[\mathbf{c}(\xi)^\top \mathbf{x}(\xi)]
\]

subject to \(\mathbf{A}(\xi)\mathbf{x}(\xi) \leq \mathbf{b}(\xi), \forall \xi \in \Xi\) such that \(\|\xi - \xi'\| \leq \rho\) for some \(\xi' \in \mathcal{S}\).

We remark that the limit as \(\rho\) tends down to zero must exist, as the optimal objective value of the above optimization problem with expanded support is monotone in \(\rho\). Note also that the expectation in the objective function has been replaced with \(\mathbb{E}_\mathcal{P}[\cdot]\), which we define here as the upper semicontinuous envelope of the objective function, i.e., \(\mathbb{E}_\mathcal{P}[f(\xi)] \equiv \mathbb{E}[\limsup_{\xi \to \xi'} f(\xi)]\). Our main result is the following.

**Theorem 5.** Suppose Assumptions 1 and 2 hold. Then, \(\mathbb{P}^\infty\)-almost surely we have

\[
J \leq \liminf_{N \to \infty} \bar{J}_N \leq \limsup_{N \to \infty} \bar{J}_N \leq \bar{J}.
\]

**Proof.** See Section EC.5. \(\square\)

Theorem 5 provides assurance that the proposed data-driven approach becomes a near-optimal approximation of an underlying stochastic multi-stage linear optimization problem. Note that Theorem 5 holds in very general cases; for example, it does not require boundedness on the space of feasible decisions, only requires the light-tail assumption on the underlying distribution, and holds when the decisions contain both continuous and integer components. Moreover, we now provide an example which illustrates that the guarantees of Theorem 5 can reflect the behavior of Problem (1) as more data is obtained. For simplicity, we discuss a single-stage problem.

**Example 3.** Consider the following single-stage stochastic optimization problem:

\[
\begin{align*}
\minimize_{x_1 \in \mathbb{Z}} & \quad x_1 \\
\text{subject to} & \quad \mathbb{P}(x_1 \geq \xi_1) = 1.
\end{align*}
\]

We assume that the true distribution is unknown, and our information consists of historical data \(\hat{\xi}_1, \ldots, \hat{\xi}_N\), which are independent and identically distributed random variables with the same distribution as \(\xi_1\). We also assume knowledge that the true support is contained in \(\Xi \equiv [0, 2]\). We consider applying Problem (1) to the above optimization problem with a robustness parameter \(\epsilon_N \equiv N^{-\frac{1}{2}}\).
To illustrate the guarantees of Theorem 5, let us consider the case where $\xi_1$ has a probability distribution defined by $\mathbb{P}(\xi_1 > \alpha) = (1 - \alpha)^k$ for some fixed $k > 0$ and $\alpha \in [0, 1]$. In this case, the true support of the underlying distribution is $S \equiv [0, 1]$ for any $k > 0$. In Section EC.6, we prove the following.

<table>
<thead>
<tr>
<th>Range of $k$</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k \in (0, 3)$</td>
<td>$\mathbb{P}^\infty \left{ J &lt; \liminf_{N \to \infty} \hat{J}<em>N = \limsup</em>{N \to \infty} \hat{J}_N = \bar{J} \right} = 1$.</td>
</tr>
<tr>
<td>$k = 3$</td>
<td>$\mathbb{P}^\infty \left{ J = \liminf_{N \to \infty} \hat{J}<em>N &lt; \limsup</em>{N \to \infty} \hat{J}_N = \bar{J} \right} = 1$.</td>
</tr>
<tr>
<td>$k \in (3, \infty)$</td>
<td>$\mathbb{P}^\infty \left{ J = \liminf_{N \to \infty} \hat{J}<em>N = \limsup</em>{N \to \infty} \hat{J}_N &lt; \bar{J} \right} = 1$.</td>
</tr>
</tbody>
</table>

This example illustrates that the bounds from Theorem 5 can hold at equality or strict inequality under different choices of $k > 0$. Moreover, this example illustrates the challenges encountered when the true support is unknown. □

6. Applications and Numerical Experiments

In this section, we demonstrate the value of the proposed methods in addressing data-driven problems of interest from the literature. Specifically, we assess the tractability and out-of-sample performance of the proposed methods in multi-stage applications where (i) uncertainty is correlated across stages, and (ii) the only knowledge regarding the uncertainty comes from historical data. In Section 6.1, we consider a multi-stage lot sizing problem for new short lifecycle products. In Section 6.2, we consider a multi-stage inventory problem with (unknown) autoregressive demands.

We compare the following approaches.

- **(SRO)** The proposed data-driven framework of Problem (1) with the finite adaptability algorithm described in Section 3. Within the partitions, static decision rules are used for binary decisions and linear decision rules are used for continuous decisions. The partitions are computed using the iterative algorithm described in Section EC.7. Unless stated otherwise, the algorithm was run with a single iteration of partitioning. We henceforth refer to this approach as sample robust optimization (SRO).

- **(Approx PCM)** The DRO method for multi-stage linear optimization proposed by Bertsimas et al. (2018c). In this approach, the ambiguity sets are constructed as distributions $\mathcal{Q} \in \mathcal{P}(\Xi)$ with the same first and second moments (mean and covariance) as the underlying distribution. As the true distribution is unknown, we estimate these moments from the historical data. The resulting DRO problem is approximated using lifted linear decision rules, as described in (Bertsimas et al. 2018c).
• (RDDP) The robust data-driven dynamic programming approach proposed by Hanasusanto and Kuhn (2013). This approach estimates the cost-to-go functions based on historical sample paths and kernel density estimation. RDDP requires both sample paths and initial state paths; we used half of the available data as these sample paths, and the other half to generate the state paths with lifted linear decision rules obtained by Approx PCM. The method also requires a robustness parameter $\gamma$, which we choose to be either $\gamma = 0$ (DDP) or $\gamma = 10$ (RDDP).

• (WDRO) The alternative approach described in Section 4.3. In this approach, we solve Problem (6) with the 1-Wasserstein ambiguity set. We approximate the optimization problem using finite adaptability from Section 3, with the reformulation described in Section EC.8. The robustness parameter and partitions are the same as SRO.

In each example, for each set of problem parameters and for each value of $N$, we generate 100 training sets. For each training set and each method, we compute the average out-of-sample cost of the resulting solutions on a testing set generated with the same parameters. The out-of-sample performance for SRO was only evaluated on valid points in the test set, i.e., points which lie within the estimated support. Therefore, we also present a feasibility measure for SRO, i.e., the percent of valid points in the test set. All algorithms were implemented in the Julia programming language (0.6.1) using JuMP (Lubin and Dunning 2015) and solved using Gurobi 7.0.1 (Gurobi Optimization 2018).

6.1. The Multi-Stage Lot-Sizing Problem

Consider a manufacturer which produces products with short lifecycles (e.g., seasonal fashion apparel or computer hardware) for a retailer. The manufacturer is obliged to quickly fulfill the retailer’s orders over each product’s lifecycle; however, the order quantities from the retailer fluctuate over time. While the demand trajectory of a new product is uncertain, the manufacturer is endowed with historical sales over the lifecycle of similar products, visualized in Figure 1. The goal is to dynamically determine the manufacturer’s production levels over a product’s lifecycle to satisfy the retailer’s orders while minimizing production and holding costs.

We formalize this setting as a multi-stage lot sizing problem$^2$. At each time period $t \in \{1, \ldots, T\}$, we (the manufacturer) start with $I_t$ units of inventory (the initial inventory is $I_1 = 0$). The demand $\xi_t \geq 0$ for the product (from the retailer) is then revealed. We must satisfy the demand in each period, and any excess inventory in each period incurs a holding cost of 65 per unit. To satisfy the demands, we have several options. (i) We can begin production of $x_{t+1} \geq 0$ units at a per-unit cost

$^2$ The problem parameters are adapted from Bertsimas and Georgiou (2015), Bertsimas and Dunning (2016).
Figure 1  Historical sales over lifecycle of similar products.

Note. Demands by retailer to the manufacturer over the lifecycle of 50 past products (simulated). Each line corresponds to a different demand trajectory $\hat{\xi}^j$ from product launch, and the $x$-axis denotes the time since the product was launched. Note that the time until obsolescence (i.e., the stage in which the demand drops to zero) is variable.

of 50, which will be available in the following period. (ii) We can produce at most two lots of 25 units to be available immediately, with a cost of 1500 for the first lot and 1875 for the second lot. We denote the binary decisions of whether to purchase the lots by $z_{t+1,1}$, $z_{t+1,2} \in \{0,1\}$. Assuming the demands are random variables, the resulting stochastic multi-stage linear optimization problem is given by

$$
\begin{aligned}
\minimize_{x, I, z} & \quad \mathbb{E}_{P} \left[ \sum_{t=1}^{T} (50x_t(\xi_{1:t-1}) + 64I_{t+1}(\xi_{1:t}) + 1500z_{t+1,1}(\xi_{1:t}) + 1875z_{t+1,2}(\xi_{1:t})) \right] \\
\text{subject to} & \quad I_{t+1}(\xi_{1:t}) = I_{t}(\xi_{1:t-1}) + x_{t}(\xi_{1:t-1}) + 25z_{t+1,1}(\xi_{1:t}) + 25z_{t+1,2}(\xi_{1:t}) - \xi_t \\
& \quad I_{t+1}(\xi_{1:t}), x_{t}(\xi_{1:t-1}) \geq 0 \\
& \quad z_{t+1,1}(\xi_{1:t}), z_{t+1,2}(\xi_{1:t}) \in \{0,1\},
\end{aligned}
$$

where the constraints must hold almost surely for each $t \in [T]$. Note that the variables $I_t$ can be eliminated from the model by recursively substituting the equality constraints.

In order to generate the historical and future data, we suppose that the demand $\xi \equiv (\xi_1, \ldots, \xi_T)$ of a new product has one of three different trajectory types: a “low” trajectory, a “medium” trajectory, and a “high” trajectory. The type of the demand trajectory is not known ahead of time, and is chosen randomly with equal probability of each type. The demand trajectories are generated according to the following process.
Figure 2  Lot Sizing Problem: Performance and Computation Time

Note. Average out-of-sample performance and solution time for the multi-stage lot sizing problem with $T = 8$ over 100 training sets of varying numbers $N$ of historical sample paths. The left graph shows the average out-of-sample cost computed on a test set of 10000 sample paths. The right graph shows the average solver time in seconds.

$$\xi_t = \begin{cases} \max\{0, 50 + \epsilon_1\}, & \text{if } t = 1, \\ \max\{0, \xi_{t-1} + \Delta^\text{type}_t + \epsilon_t\}, & \text{if } t > 1, \xi_{t-1} > 0, \\ 0, & \text{if } t > 1, \xi_{t-1} = 0; \end{cases}$$

where $\epsilon_1, \ldots, \epsilon_T$ are independently and identically distributed Gaussian random variables with mean zero and standard deviation of 7 units.

$\Delta^\text{type}_t = \begin{cases} 20 - 20(t - 1), & \text{if type = low}, \\ 38 - 16(t - 1), & \text{if type = medium}, \\ 35 - 10(t - 1), & \text{if type = high}, \end{cases}$

We address Problem (9) using SRO, Approx PCM, and WDRO\textsuperscript{3}. The methods SRO and WDRO use the robustness parameter of $\epsilon_N = 45N^{-1/(T+1)}$. We observe that Approx PCM and WDRO require a bounded $\Xi$ in order to be feasible. Thus, to enable comparison of the different methods, we set $\Xi$ to be the hyperrectangle $\Xi = [0, 150]^T$.

In Figure 2, we present the out-of-sample cost and solutions times for the various methods with $T = 8$. The results demonstrate that SRO produces decision rules with an average cost that significantly outperforms the alternative methods. In particular, the average cost of decision rules from SRO is lower than that from WDRO by 18.5\% (when $N = 10$) and 32\% (when $N = 1000$). Approx PCM, while conservative, takes less than a second to solve; in contrast, SRO may take up to five minutes to solve (for $N = 1000$) but obtains a much lower cost. In Table 1, we demonstrate that, under our choice of robustness parameter, SRO produces decision rules which are feasible with high probability.

\textsuperscript{3} RDDP was not applied for this example due to the high computational cost. Indeed, running RDDP for the inventory problem in Section 6.2 takes over 1000 seconds when $N = 100$, and the lot sizing problem additionally requires integer variables.
Table 1  Lot sizing problem with $T = 8$ - SRO Feasibility.

<table>
<thead>
<tr>
<th>N</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
<th>600</th>
<th>800</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>95.63</td>
<td>99.46</td>
<td>99.79</td>
<td>99.88</td>
<td>99.91</td>
<td>99.93</td>
<td>99.93</td>
<td>99.94</td>
<td>99.95</td>
</tr>
<tr>
<td>Min</td>
<td>67.18</td>
<td>96.67</td>
<td>98.57</td>
<td>99.43</td>
<td>99.74</td>
<td>99.80</td>
<td>99.80</td>
<td>99.84</td>
<td>99.87</td>
</tr>
</tbody>
</table>

Feasibility (%) of SRO for the multi-stage lot sizing with $T = 8$ computed over 100 training sets. Mean percentage of valid points from the 10,000 test set points that fall within the estimated support.

Figure 3  Lot Sizing Problem: Optimality Gap for $T = 3$

Note. Average optimality gap and running time for the lot-sizing problem with $T = 3$ with a training set of $N = 1,000$ sample paths. The left graph shows the optimality gap, which is the difference (in percentages) between the average out-of-sample cost computed on the test set of 10,000 sample paths and an estimate of the optimal cost of Problem (9). The estimated optimal cost was obtained by SAA, which involves conditionally sampling from the true distribution (see Shapiro et al. (2009)). We constructed 100 scenario trees, each with a $40^3$ points (each branch of the scenario tree with 40 children), and took the average over their objective value as our approximation. The right graph shows the average total solution time in seconds. The results for SRO and WDRO are performed for varying numbers of iterations.

To estimate the optimality gap of the various methods, we performed additional experiments for a variant of the lot sizing problem where $T = 3$. In this case, we can compute a lower bound on the optimal cost of Problem (9) using SAA. Figure 3 shows the average optimality gap and running times of the compared methods for $N = 1000$ and multiple iterations of the partitioning algorithm. The results show that SRO finds decision rules with an average out-of-sample that is within 50% of the lower bound with more iterations of the partitioning algorithm.
6.2. Multi-Stage Inventory Control

Many previous approaches have been proposed to address multi-stage inventory control problems with autoregressive demand models. In this section, we apply the proposed data-driven framework and compare against alternatives.

In this problem, we have a finite planning horizon of length $T$. At the beginning of the $t$-th time period, we have $I_t$ units in inventory and we order $x_t \in [0, \bar{x}_t]$ units of product with zero lead time at a cost of $c_t$ per unit. Demand $\xi_t \geq 0$ is then revealed, and the inventory $I_{t+1}$ in the following period is set to $I_t + x_t - \xi_t$; hence, the inventory is fully backlogged. We incur a holding cost of $h_t \max\{I_t + x_t - \xi_t, 0\}$ and a backorder cost of $b_t \max\{\xi_t - I_t - x_t, 0\}$. We start on period $t = 1$ with $I_1 = 0$ units of inventory. On the final period $t = T$, the salvage value of leftover inventory can be accounted for with $h_T$ and $b_T$. Our goal is to minimize the average cost per planning horizon, given by the following optimization problem.

\[
\begin{align*}
\text{minimize} \quad & \mathbb{E}_p \left[ \sum_{t=1}^{T} (c_t x_t(\xi_{1:t-1}) + y_{t+1}(\xi_{1:t})) \right] \\
\text{subject to} \quad & I_{t+1}(\xi_{1:t}) = I_t(\xi_{1:t-1}) + x_t(\xi_{1:t-1}) - \xi_t \\
& y_{t+1}(\xi_{1:t}) \geq h_t I_{t+1}(\xi_{1:t}) \\
& y_{t+1}(\xi_{1:t}) \geq -b_t I_{t+1}(\xi_{1:t}) \\
& 0 \leq x_t(\xi_{1:t-1}) \leq \bar{x}_t z_t(\xi_{1:t-1}).
\end{align*}
\]

where $y_{t+1}$ is the cost incurred by the inventory at time $t + 1$, and the constraints must hold almost surely for every $t \in \{1, \ldots, T\}$. Note that the inventory variables can be eliminated from this formulation by recursively applying the equality constraints.

As in previous literature (e.g., Graves (1999), See and Sim (2010)), we consider the case where the underlying demand process is a non-stationary stochastic process of the form

\[\xi_t = \varsigma_t + \alpha \varsigma_{t-1} + \cdots + \alpha \varsigma_1 + \mu,\]

where $\varsigma_1, \ldots, \varsigma_T$ are independent random variables distributed uniformly over $[-\varsigma, \varsigma]$. We assume that the the specific values of $\alpha$, $\mu$ and $\varsigma$ are unknown, and assume that the support is known to lie in the nonnegative orthant. For computational experiments, we use the problem parameters from Bertsimas et al. (2018c): $\bar{x}_t = 260$, $c_t = 0.1$, $h_t = 0.02$ for all $t \in [T]$, $b_t = 0.2$ for all $t \in [T-1]$, $b_T = 2$, and $\alpha \in \{0, 0.25, 0.5\}$. For $T = 5$, we used $\mu = 200$ and $\varsigma = 20$, and for $T = 10$ we used $\mu = 200$ and $\varsigma = 10$.

In Tables 2 and 3, we display the out-of-sample performance of the various methods and the feasibility of SRO, respectively. The results demonstrate that SRO has a superior out-of-sample
### Table 2  Multi-stage inventory problem out-of-sample performance.

<table>
<thead>
<tr>
<th>$T$</th>
<th>$\alpha$</th>
<th>Method</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0</td>
<td>SRO</td>
<td>110.5(1.8)</td>
<td>108.2(1.0)</td>
<td>107.2(0.7)</td>
<td><strong>106.4(0.6)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approx PCM</td>
<td>109.3(1.1)</td>
<td>108.7(0.5)</td>
<td>108.6(0.3)</td>
<td>108.5(0.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DDP</td>
<td>2283.8(349.2)</td>
<td>1174.7(870.6)</td>
<td>501.3(491.2)</td>
<td>203.2(202.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RDDP</td>
<td>2283.1(350.6)</td>
<td>1176.4(873.4)</td>
<td>501.3(488.9)</td>
<td>201.7(199.7)</td>
</tr>
<tr>
<td>0.25</td>
<td></td>
<td>SRO</td>
<td>110.3(1.7)</td>
<td>108.2(0.9)</td>
<td>107.7(0.7)</td>
<td><strong>107.4(0.6)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approx PCM</td>
<td>113.5(1.9)</td>
<td>112.7(0.7)</td>
<td>112.5(0.4)</td>
<td>112.4(0.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DDP</td>
<td>2278.4(400.7)</td>
<td>1297.6(918.9)</td>
<td>617.7(637.2)</td>
<td>194.9(217.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RDDP</td>
<td>2277.9(410.0)</td>
<td>1294.2(919.8)</td>
<td>613.6(635.0)</td>
<td>196.1(216.3)</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td>SRO</td>
<td>110.9(2.3)</td>
<td>109.2(1.4)</td>
<td>108.7(1.0)</td>
<td><strong>108.4(0.8)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approx PCM</td>
<td>118.2(2.5)</td>
<td>117.1(0.9)</td>
<td>116.8(0.6)</td>
<td>116.6(0.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DDP</td>
<td>2260.4(458.8)</td>
<td>1366.9(954.3)</td>
<td>694.7(728.7)</td>
<td>227.5(282.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RDDP</td>
<td>2260.0(459.3)</td>
<td>1364.2(955.0)</td>
<td>691.2(726.1)</td>
<td>229.2(281.2)</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>SRO</td>
<td>211.6(1.5)</td>
<td>209.8(1.1)</td>
<td>208.4(0.8)</td>
<td><strong>207.3(0.6)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approx PCM</td>
<td>208.3(1.1)</td>
<td>207.7(0.3)</td>
<td>207.5(0.2)</td>
<td>207.5(0.2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DDP</td>
<td>5182.1(1253.6)</td>
<td>2885.7(1872.4)</td>
<td>1209.7(1181.6)</td>
<td>387.3(283.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RDDP</td>
<td>5181.0(1255.0)</td>
<td>2880.7(1872.9)</td>
<td>1227.3(1197.9)</td>
<td>396.3(296.3)</td>
</tr>
<tr>
<td>0.25</td>
<td></td>
<td>SRO</td>
<td>210.7(1.5)</td>
<td>209.0(0.8)</td>
<td>208.1(0.6)</td>
<td><strong>207.6(0.6)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approx PCM</td>
<td>214.8(2.4)</td>
<td>213.7(0.8)</td>
<td>213.4(0.5)</td>
<td>213.3(0.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DDP</td>
<td>5295.4(1307.4)</td>
<td>3048.7(2103.9)</td>
<td>1334.7(1428.1)</td>
<td>362.4(377.4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RDDP</td>
<td>5289.7(1314.5)</td>
<td>3046.0(2105.1)</td>
<td>1332.7(1438.0)</td>
<td>392.4(383.4)</td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td>SRO</td>
<td>211.9(2.9)</td>
<td>210.4(1.5)</td>
<td>209.6(1.0)</td>
<td><strong>209.0(0.8)</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approx PCM</td>
<td>222.6(3.5)</td>
<td>220.8(1.3)</td>
<td>220.4(0.8)</td>
<td>220.1(0.8)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DDP</td>
<td>5304.8(1331.1)</td>
<td>3218.9(2244.3)</td>
<td>1402.3(1686.8)</td>
<td>370.1(618.9)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RDDP</td>
<td>5298.7(1336.3)</td>
<td>3216.0(2244.8)</td>
<td>1481.5(1740.2)</td>
<td>412.2(643.3)</td>
</tr>
</tbody>
</table>

Mean (standard deviation) of out-of-sample cost for the multi-stage inventory problem over 100 training sets. The average out-of-sample cost for each training set is computed on a test set of 1,000 sample paths. Note that the out-of-sample cost was computed as $\sum_{t=1}^{T} (c_t x_t + \max \{ h_t I_{t+1} - b_t I_{t+1} \})$, as the decision rules may result in $y_t$s which are conservative upper bounds.

### Table 3  Multi-stage inventory problem - SRO Feasibility.

<table>
<thead>
<tr>
<th>$T$</th>
<th>$\alpha$</th>
<th>Size of Training Set (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>99.71(1.66)</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>99.40(1.88)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>98.90(2.39)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>100.00(0.00)</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>99.92(0.21)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>99.34(1.21)</td>
</tr>
</tbody>
</table>

Feasibility (%) for SRO in the multi-stage inventory problem. Mean (standard deviation) of the percentage of valid points from a test set of 10,000 sample paths that fell within the estimated support. Computed over 100 training sets.

average performance to alternative approaches for most problem parameters. In comparison to Approx PCM, SRO achieves a better out-of-sample performance for almost all problem parameters. This improvement comes with a greater computational cost, which is illustrated in Figure 4. While
Figure 4 Multi-stage inventory problem running times.

Note. Average running times for the multi-stage inventory control problem with five stages ($T = 5$) and no correlation between time steps ($\alpha = 0$). The graph shows the mean value of the running times over 100 training sets for different training set sizes ($N$).

Approx PCM takes at most 5 seconds to solve for any size training set and any value of $\alpha$, SRO’s running times increase with both $N$ and $\alpha$, reaching a maximum of two minutes for $T = 10$, $N = 100$, and $\alpha = 0.5$.

In general, we did not observe a significant difference between the performance of DDP and RDDP. Although DDP and RDDP performed poorly on average, these methods have a high standard deviation and thus outperformed SRO on some training instances. SRO is more computationally tractable than DDP and RDDP, as demonstrated in Figure 4. For example, when $T = 5$ and $\alpha = 0$, DDP and RDDP have an average computation time that was $100x$ that of SRO. We note that one could partially parallelize DDP and RDDP, which might help improve the running times of these approaches.

7. Conclusion

In this work, we presented the first data-driven approach for multi-stage linear optimization problems, including those with integer decisions, that combines practical tractability with nonparametric convergence guarantees. In particular, the proposed method is conceptually simple and addresses multi-stage problems where uncertainty is arbitrarily correlated across stages. The practical value of the proposed approach was illustrated by computational examples motivated by real-world applications, demonstrating that the proposed data-driven approach produced high-quality solutions.
in reasonable computation time. Apart from these contributions, the paper also presented general results on distributionally robust optimization using the $\infty$-Wasserstein ambiguity set, which may have implications in areas such as statistical learning theory.

References


Electronic Companion

EC.1. Proof of Lemma 1 and Theorem 1 from Section 4.2

For convenience, we restate Lemma 1.

**Lemma 1** The $\infty$-Wasserstein ambiguity set is equivalent to

\[
\left\{ \frac{1}{N} \sum_{j=1}^{N} Q_j \left( \|\xi - \hat{\xi}^j\| \leq \epsilon_N \right) = 1 \text{ for each } j \in [N], \right. \\
\left. Q_1, \ldots, Q_N \in \mathcal{P}(\Xi) \right\}.
\]

**Proof.** By the definition of the extended-valued $\infty$-Wasserstein distance from Section 4.1, the $\infty$-Wasserstein ambiguity set of interest has the form

\[
\left\{ Q \in \mathcal{P}(\Xi) : d_{\infty}(Q, \hat{P}_N) \leq \epsilon_N \right\} = \left\{ Q \in \mathcal{P}(\Xi) : \right. \\
\Pi \left( \|\xi - \xi'\| \leq \epsilon_N \right) = 1, \text{ and } \\
\Pi \text{ is a joint distribution of } \xi \text{ and } \xi' \\
\text{with marginals } Q \text{ and } \hat{P}_N, \text{ respectively} \right\}. \tag{EC.1}
\]

Let $\hat{\xi}^1, \ldots, \hat{\xi}^K$ be the distinct values among $\hat{\xi}^1, \ldots, \hat{\xi}^N$, and let $I_1, \ldots, I_K$ be index sets where

\[
I_k = \{ j \in [N] : \hat{\xi}^j = \hat{\xi}^k \}.
\]

For any joint distribution $\Pi$ that satisfies the constraints in the ambiguity set on Line (EC.1), let $Q_k$ be the conditional distribution of $\xi$ given $\xi' = \hat{\xi}^k$. Then, for every Borel set $A \subseteq \Xi$,

\[
Q(\xi \in A) = \Pi \left( (\xi, \xi') \in A \times \Xi \right) = \sum_{k=1}^{K} \Pi \left( \xi' = \hat{\xi}^k \right) \hat{P}_N(\xi' = \hat{\xi}^k) = \sum_{k=1}^{K} Q_k(\xi \in A) \frac{|I_k|}{N}.
\]

The first equality follows because $\Pi$ is a joint distribution of $\xi$ and $\xi'$ with marginals $Q$ and $\hat{P}_N$, respectively. The second equality follows from the law of total probability. The final equality follows from the definition of the conditional distribution $Q_j$ and the empirical distribution $\hat{P}_N$. Since this reasoning holds for every Borel set, we have shown that $Q = \sum_{k=1}^{K} \frac{|I_k|}{N} Q_k$. Furthermore, following similar reasoning as above, we observe that

\[
\Pi \left( \|\xi - \xi'\| \leq \epsilon_N \right) = \sum_{k=1}^{K} \Pi \left( \|\xi - \xi'\| \leq \epsilon_N | \xi' = \hat{\xi}^k \right) \hat{P}_N(\xi' = \hat{\xi}^k) = \sum_{k=1}^{K} Q_k(\|\xi - \hat{\xi}^k\| \leq \epsilon_N) \frac{|I_k|}{N}.
\]

Therefore, combining the above results, the ambiguity set from Line (EC.1) can be rewritten as
Lemma 1 implies that the infinite-Wasserstein ambiguity set can be decomposed into separate distributions, each having a support that is contained in $Q_j$. The first equality follows because each $Q_k$ satisfies $Q_k(\|\xi - \bar{\xi}^k\| \leq \epsilon_N) = 1$, for each $k \in [K]$. Moreover, it similarly follows from the definition of the infinite-Wasserstein ambiguity set.

For convenience, we also restate Theorem 1.

**Theorem 1** Problem (1) is equivalent to Problem (6) using the infinite-Wasserstein ambiguity set.

**Proof.** It follows from Lemma 1 that the infinite-Wasserstein ambiguity set can be decomposed into separate distributions, each having a support that is contained in $\{\xi \in \Xi : \|\xi - \bar{\xi}^j\| \leq \epsilon_N\}$ for each $j \in [N]$. Of course, these sets are exactly equal to the uncertainty sets from Section 2.3, and thus Lemma 1 implies that the infinite-Wasserstein ambiguity set is equivalent to

$$\left\{ \frac{1}{N} \sum_{j=1}^{N} Q_j : Q_j \in \mathcal{P}(U^j_N) \text{ for each } j \in [N] \right\}.$$

Therefore, when $A$ is the infinite-Wasserstein ambiguity set, we readily observe that

$$\sup_{Q \in A} \mathbb{E}_Q [c(\xi)^\top x(\xi)] = \frac{1}{N} \sum_{j=1}^{N} \sup_{Q_j \in \mathcal{P}(U^j_N)} \mathbb{E}_Q [c(\xi)^\top x(\xi)] = \frac{1}{N} \sum_{j=1}^{N} \sup_{\xi \in U^j_N} c(\xi)^\top x(\xi).$$

Moreover, it similarly follows from the definition of the infinite-Wasserstein ambiguity set that

$$Q(A(\xi)x(\xi) \leq b(\xi)) = 1, \forall Q \in A \iff \frac{1}{N} \sum_{j=1}^{N} Q_j(A(\xi)x(\xi) \leq b(\xi)) = 1, \forall Q_j \in \mathcal{P}(U^j_N), j \in [N]$$

$$\iff Q_j(A(\xi)x(\xi) \leq b(\xi)) = 1, \forall Q_j \in \mathcal{P}(U^j_N), j \in [N]$$

$$\iff A(\zeta)x(\zeta) \leq b(\zeta), \forall \zeta \in U^j_N, j \in [N].$$

This concludes the equivalence.

EC.2. Proof of Theorem 2 from Section 4.2

For convenience, we restate Theorem 2.
**Theorem 2** Let $A$ be the $p$-Wasserstein ambiguity set with robustness parameter $\epsilon_N > 0$ and $p \in [1, \infty)$. Then, the constraints from Problem (6), i.e.,

$$Q(A(\xi)x(\xi) \leq b(\xi)) = 1, \quad \forall Q \in A,$$

are satisfied if and only if

$$A(\zeta)x(\zeta) \leq b(\zeta) \quad \forall \zeta \in \Xi.$$

**Proof.** Choose any $\bar{\xi} \in \Xi$ such that $\bar{\xi} \neq \hat{\xi}^j$ for each $j \in [N]$. For any $\lambda \in [0, 1]$, define

$$Q^\lambda \equiv (1 - \lambda)\hat{P}_N + \lambda \delta_{\bar{\xi}}.$$

where $\hat{P}_N$ is the empirical distribution of the sample paths and $\delta_{\bar{\xi}}$ is the Dirac delta measure on $\xi$. By the definition of the $p$-Wasserstein distance,

$$d_p\left(\hat{P}_N, Q^\lambda\right) = \inf \left\{ \left( \int_{\Xi \times \Xi} \|\xi - \xi'\|^p d\Pi(\xi, \xi') \right)^{\frac{1}{p}} : \Pi \text{ is a joint distribution of } \xi \text{ and } \xi' \right\}$$

with marginals $\hat{P}_N$ and $Q^\lambda$, respectively.

We now consider a feasible (but possibly suboptimal) joint distribution $\bar{\Pi}$ for the above optimization problem in which $\xi' \sim Q^\lambda$, $\xi'' \sim \hat{P}_N$, and

$$\xi = \begin{cases} \xi', & \text{if } \xi' = \hat{\xi}^j \text{ for some } j \in [N], \\ \xi'', & \text{otherwise.} \end{cases}$$

Under the assumption that $\lambda \in [0, 1]$, it is readily verified that the marginal distributions of $\xi$ and $\xi'$ are $\hat{P}_N$ and $Q^\lambda$, respectively. Then,

$$d_p\left(\hat{P}_N, Q^\lambda\right) \leq \left( \int_{\Xi \times \Xi} \|\xi - \xi'\|^p d\bar{\Pi}(\xi, \xi') \right)^{\frac{1}{p}}$$

$$= \left( \int_{\Xi \times \Xi} \|\xi - \xi'\|^p 1 \{\xi' = \bar{\xi}\} d\bar{\Pi}(\xi, \xi') + \int_{\Xi \times \Xi} \|\xi - \xi'\|^p 1 \{\xi' \neq \bar{\xi}\} d\bar{\Pi}(\xi, \xi') \right)^{\frac{1}{p}}$$

$$= \left( \frac{1}{N} \sum_{j=1}^{N} \lambda \|\hat{\xi}^j - \bar{\xi}\|^p \right)^{\frac{1}{p}}.$$

The inequality follows since $\bar{\Pi}$ is feasible but possibly suboptimal joint distribution. The first equality follows from splitting the integral into two cases, and observing that the second case equals zero since $\xi = \xi'$ whenever $\xi' \neq \bar{\xi}$. The final equality follows because $\xi = \xi''$ whenever $\xi' = \bar{\xi}$, and
ξ" is distributed uniformly over the empirical measure. Thus, we have shown that $Q^\lambda$ is contained in the $p$-Wasserstein ambiguity set when

$$\tilde{\lambda} = \min \left\{ 1, \frac{\epsilon_P}{\frac{1}{N} \sum_{j=1}^{N} \|\hat{\xi} - \bar{\xi}\|^p} \right\}.$$ 

Suppose $\epsilon_N > 0$. Then, for any decision rule that satisfies the constraints of Problem (6), it must follow that

$$Q^\bar{\lambda}(A(\xi)x(\xi) \leq b(\xi)) = 1.$$ 

Since $\epsilon_N > 0$ implies that $\tilde{\lambda} > 0$, the above constraint implies that $A(\bar{\xi})x(\bar{\xi}) \leq b(\bar{\xi})$ and $A(\hat{\xi}^j)x(\hat{\xi}^j) \leq b(\hat{\xi}^j)$ for each $j \in [N]$. Since $\bar{\xi} \in \Xi$ was picked arbitrarily, we conclude that the constraints of Problem (6) are satisfied only if

$$A(\zeta)x(\zeta) \leq b(\zeta), \quad \forall \zeta \in \Xi.$$ 

Note that the opposite follows directly from the fact that $A \subseteq P(\Xi)$. □

**EC.3. Proof of Theorem 3 from Section 5.2**

The methodology of the following proof follows similar reasoning to Devroye and Wise (1980) and Baillé et al. (2000), which is adapted to the light-tail assumption of this paper. The theorem is restated here for convenience.

**Theorem 3.** Suppose Assumptions 1 and 2 hold. Then

$$\left( \frac{N^{\frac{1}{d+1}}}{(\log N)^{d+1}} \right) \mathbb{P}(\xi \not\in S_N) \to 0$$ 

as $N \to \infty$, $\mathbb{P}^\infty$-almost surely.

**Proof.** Choose any arbitrary $\eta > 0$, and let $R_N = N^{\frac{1}{d+1}}(\log N)^{-(d+1)}$. Moreover, let $a > 1$ be a fixed constant such that $b \equiv \mathbb{E}[\exp(\|\xi\|^a)] < \infty$ (the existence of $a$ and $b$ follows from Assumption 1). Define

$$A_N \equiv \left\{ \zeta \in \mathbb{R}^d : \|\zeta\| \leq (\log N)^{\frac{a+1}{2a}} \right\}.$$ 

We begin by showing that $R_N\mathbb{P}(\xi \not\in A_N) \leq \eta$ for all sufficiently large $N$. Indeed,

$$R_N\mathbb{P}(\xi \not\in A_N) = R_N\mathbb{P}\left(\|\xi\| > (\log N)^{\frac{a+1}{2a}}\right) = R_N\mathbb{P}\left(\exp(\|\xi\|^a) > \exp((\log N)^{\frac{a+1}{2a}})\right) \leq \frac{bR_N}{\exp((\log N)^{\frac{a+1}{2a}})} \leq \eta.$$
The first inequality follows from Markov’s inequality and the second inequality holds for all sufficiently large $N$ since $a > 1$.

Next, define

$$
\alpha_N \equiv \frac{\eta}{2^d (\log N)^{\frac{a+1}{2a}}} R_N ; \quad B_N \equiv \{ \zeta \in \mathbb{R}^d : \mathbb{P}(\|\xi - \zeta\| \leq \epsilon_N) > \alpha_N \epsilon_N^d \}.
$$

We now show that $R_N \mathbb{P}(\xi \notin B_N) \leq 2\eta$ for all sufficiently large $N$. Indeed, for all sufficiently large $N$,

$$
R_N \mathbb{P}(\xi \notin B_N) = R_N \mathbb{P}(\xi \in A_N, \xi \notin B_N) + R_N \mathbb{P}(\xi \notin A_N, \xi \in B_N)
$$

$$
\leq R_N \mathbb{P}(\xi \in A_N, \xi \notin B_N) + R_N \mathbb{P}(\xi \notin A_N)
$$

$$
\leq R_N \mathbb{P}(\xi \in A_N, \xi \notin B_N) + \eta,
$$

where the first equality follows from the law of total probability and the second inequality follows because $R_N \mathbb{P}(\xi \notin A_N) \leq \eta$ for all $N$ sufficiently large. Now, choose points $\zeta^1, \ldots, \zeta^{K_N} \in A_N$ such that $\min_{j \in [K_N]} \|\zeta - \zeta^j\| \leq \frac{\epsilon_N}{2}$ for all $\zeta \in A_N$. For example, one can place the points on a grid overlaying $A_N$. It follows from Verger-Gaugry (2005) that this can be accomplished with a number of points $K_N$ which satisfies

$$
K_N \leq \phi \left( \frac{(\log N)^{\frac{a+1}{2a}}}{\epsilon_N} \right)^d.
$$

where $\phi > 0$ is a constant which depends only on $d$. Then, continuing the previous argument,

$$
R_N \mathbb{P}(\xi \notin B_N) \leq R_N \mathbb{P}(\xi \in A_N, \xi \notin B_N) + \eta \leq R_N \sum_{j=1}^{K_N} \mathbb{P}(\|\xi - \zeta^j\| \leq \frac{\epsilon_N}{2}, \xi \notin B_N) + \eta,
$$

(EC.2)

where the second inequality follows from the union bound. We now develop an upper bound on the above probability. To do so, we consider two cases. First, suppose there exists a realization $\zeta \notin B_N$ such that $\|\zeta - \zeta^j\| \leq \frac{\epsilon_N}{2}$. Then,

$$
\mathbb{P}(\|\xi - \zeta^j\| \leq \frac{\epsilon_N}{2}, \xi \notin B_N) \leq \mathbb{P}(\|\xi - \zeta\| \leq \frac{\epsilon_N}{2}) \leq \mathbb{P}(\|\xi - \zeta\| \leq \epsilon_N) \leq \alpha_N \epsilon_N^d,
$$

where the second inequality follows because $\|\xi - \zeta\| \leq \|\xi - \zeta^j\| + \|\zeta^j - \zeta\| \leq \epsilon_N$ whenever $\|\xi - \zeta^j\| \leq \frac{\epsilon_N}{2}$, and the third inequality follows because $\zeta \notin B_N$. Second, suppose there does not exist a realization $\zeta \notin B_N$ such that $\|\zeta - \zeta^j\| \leq \frac{\epsilon_N}{2}$. Then,

$$
\mathbb{P}(\|\xi - \zeta^j\| \leq \frac{\epsilon_N}{2}, \xi \notin B_N) = 0.
$$
In both cases, we have shown that $P(\|\xi - \zeta_j\| \leq \frac{\alpha N}{d}, \xi \not\in B_N) \leq \alpha N \epsilon^d_N$. Therefore, we obtain the following upper bound on Line (EC.2) for all sufficiently large $N$:

$$R_N P(\xi \not\in B_N) \leq R_N K_N \alpha N \epsilon^d_N + \eta \leq \phi(\log N) \frac{d(a+1)}{2d} R_N \alpha N + \eta \leq 2 \eta. \quad \text{(EC.3)}$$

The second inequality follows by the definition of $K_N$, and the third inequality follows for all sufficiently large $N$ by the definition of $R_N$. Thus, for all sufficiently large $N$,

$$R_N P(\xi \not\in S_N) = R_N P(\xi \not\in S_N, \xi \in B_N) + R_N P(\xi \not\in S_N, \xi \not\in B_N) \leq R_N P(\xi \not\in S_N, \xi \in B_N) + 2 \eta. \quad \text{(EC.4)}$$

where the equality follows from the law of total probability, and the inequality follows from Line (EC.3). Therefore,

$$P^N (R_N P(\xi \not\in S_N) > 3 \eta) \leq P^N (R_N P(\xi \not\in S_N, \xi \in B_N) > \eta) \leq \eta^{-1} R_N E_{\mathcal{P}^N} \left[ P \left( \xi \not\in S_N, \xi \in B_N \mid \hat{\xi}_1, \ldots, \hat{\xi}_N \right) \right] \quad \text{(EC.5)}$$

$$= \eta^{-1} R_N E_{\mathcal{P}} \left[ P^N \left( \xi \not\in S_N, \xi \in B_N \mid \xi \right) \right] \quad \text{(EC.6)}$$

$$= \eta^{-1} R_N E_{\mathcal{P}} \left[ P^N \left( \|\xi - \hat{\xi}_1\| > \epsilon_N, \ldots, \|\xi - \hat{\xi}_N\| > \epsilon_N, \xi \in B_N \mid \xi \right) \right] \quad \text{(EC.7)}$$

$$= \eta^{-1} R_N E_{\mathcal{P}} \left[ P \left( \|\xi - \xi^e\| > \epsilon_N, \xi \in B_N \mid \xi \right) \right]^N \quad \text{(EC.8)}$$

$$\leq \eta^{-1} R_N (1 - \alpha N \epsilon^d_N)^N \quad \text{(EC.9)}$$

$$\leq \eta^{-1} R_N \exp \left( -N \alpha N \epsilon^d_N \right) \quad \text{(EC.10)}$$

$$= \eta^{-1} R_N \exp \left( -\eta 2^{-d(\log N)^{1+\rho}} \right) \quad \text{(EC.11)}$$

$$\leq \eta^{-1} R_N \exp \left( -\eta 2^{-d(\log N)^{1+\rho}} \right) \quad \text{(EC.12)}$$

where $\rho \equiv \frac{d(a-1)}{2d} > 0$. Line (EC.5) follows from Line (EC.4), Line (EC.6) follows from Markov’s inequality, and Line (EC.7) follows from the law of iterated expectation. Line (EC.8) follows from the definition of $S_N$, and Line (EC.9) follows because, conditional on $\xi$, the random variables $\|\xi - \hat{\xi}_1\|, \ldots, \|\xi - \hat{\xi}_N\|$ are independent. Line (EC.10) follows from the definition of $B_N$, Line (EC.11) follows from the mean value theorem, and Line (EC.12) follows from the definition of $\alpha_N$. Therefore, we have shown that

$$\sum_{N=1}^{\infty} P^N (R_N P(\xi \not\in S_N) > 3 \eta) < \infty$$

for every $\eta > 0$, and thus the Borel-Cantelli lemma implies that $R_N P(\xi \not\in S_N) \to 0$ as $N \to \infty$, $P^\infty$-almost surely. □
EC.4. Proof of Theorem 4 from Section 5.2

In Section EC.4.1, we present an intermediary result on the gap between DRO using the 1- and ∞-Wasserstein ambiguity sets. In Section EC.4.2, we present a generalization of Theorem 4. Some intermediary and technical results are relegated to Section EC.4.3.

EC.4.1. On the gap between the 1-Wasserstein and ∞-Wasserstein ambiguity sets

In this section, we show that DRO using the ∞-Wasserstein ambiguity set is, under certain conditions, nearly an upper bound on DRO using the 1-Wasserstein ambiguity set.

We begin with a slight generalization of the 1-Wasserstein ambiguity set \((p \in [1, \infty])\) from Section 4:

\[
\mathcal{W}_p(\xi^1, \ldots, \xi^N; \theta, Z) = \left\{ Q \in \mathcal{P}(Z) : d_p\left(Q, \hat{P}_N\right) \leq \theta \right\},
\]

where \(\hat{P}_N\) is the empirical distribution of the historical data \(\xi^1, \ldots, \xi^N \subseteq \mathbb{R}^d, Z \subseteq \mathbb{R}^d\) is any set that contains the historical data, and \(\theta \geq 0\) is the robustness parameter. Note that \(\mathcal{Z}\) need not be equal to \(\Xi\). We begin with an intermediary result, in which we establish a bound for the case where there is a single data point \((N = 1)\).

**Lemma EC.1.** Let \(Z \subseteq \mathbb{R}^d, f : Z \rightarrow \mathbb{R}\) be measurable, and \(\hat{\xi} \in Z\). If \(\theta_2 \geq 2\theta_1 \geq 0\), then

\[
\sup_{Q \in \mathcal{W}_1(\xi; \theta_1, Z)} \mathbb{E}_Q[f(\xi)] \leq \sup_{Q \in \mathcal{W}_\infty(\xi; \theta_2, Z)} \mathbb{E}_Q[f(\xi)] + \frac{2\theta_1}{\theta_2} \sup_{\xi \in Z} |f(\xi)|. \tag{EC.13}
\]

**Proof.** First, it follows from the definition of the 1-Wasserstein distance (see Section 4) that

\[
\sup_{Q \in \mathcal{W}_1(\xi; \theta_1, Z)} \mathbb{E}_Q[f(\xi)] = \sup_{Q \in \mathcal{P}(Z)} \mathbb{E}_Q \left[ f(\xi) \right] \quad \text{subject to} \quad \mathbb{E}_Q \left[ f(\xi) \right] \leq \theta_1. \tag{EC.14}
\]

We now apply the Richter-Rogonsinski Theorem (see Theorem 7.32 and Proposition 6.40 of Shapiro et al. (2009)), which says that a DRO problem with \(m\) moment constraints is equivalent to optimizing a weighted average of \(m + 1\) points results. Thus, it follows from Line (EC.14) that

\[
\sup_{Q \in \mathcal{W}_1(\xi; \theta_1, Z)} \mathbb{E}_Q[f(\xi)] = \left\{ \begin{array}{ll}
\sup_{\xi^1, \xi^2 \subseteq \mathbb{R}^d, \lambda \in [0,1]} & \lambda f(\xi^1) + (1 - \lambda) f(\xi^2) \\
\text{subject to} & \lambda \left\| \xi^1 - \hat{\xi} \right\| + (1 - \lambda) \left\| \xi^2 - \hat{\xi} \right\| \leq \theta_1 \end{array} \right\}. \tag{EC.15}
\]

Let us assume from this point onward that \(\sup_{\xi \in Z} |f(\xi)| < \infty;\) indeed, if \(\sup_{\xi \in Z} |f(\xi)| = \infty,\) then the inequality in Line (EC.13) would trivially hold since the right-hand side would equal infinity. Then, it follows from Lemma 1 and Line (EC.15) that

\[
\sup_{Q \in \mathcal{W}_1(\xi; \theta_1, Z)} \mathbb{E}_Q[f(\xi)] = \sup_{0 \leq \lambda \leq 1} \left\{ \lambda \left( \sup_{Q \in \mathcal{W}_\infty(\xi; \theta_1, Z)} \mathbb{E}_Q[f(\xi)] \right) + (1 - \lambda) \left( \sup_{Q \in \mathcal{W}_\infty(\xi; \theta_1, Z)} \mathbb{E}_Q[f(\xi)] \right) \right\}. \tag{EC.16}
\]
We observe that the outer-most optimization problem in Line (EC.16) is symmetric with respect to \( \lambda \), in the sense that \( \lambda \) can be restricted to \([0, \frac{1}{2}] \) or \([\frac{1}{2}, 1] \) without loss of generality. Moreover, under the assumption that \( \theta_2 \geq 2\theta_1 \), the interval \([0, 1 - \frac{\theta_1}{\theta_2}] \) is a superset of the interval \([0, \frac{1}{2}] \). Combining these arguments, we conclude from Line (EC.16) that

\[
\sup_{Q \in W_1(\xi; \theta_1, z)} E_Q [f(\xi)] = \sup_{0 \leq \lambda \leq 1 - \frac{\theta_1}{\theta_2}} \left\{ \lambda \left( \sup_{Q \in W_\infty(\xi; \frac{\theta_1}{2}, z)} E_Q [f(\xi)] \right) + (1 - \lambda) \left( \sup_{Q \in W_\infty(\xi; \frac{\theta_1}{2}, z)} E_Q [f(\xi)] \right) \right\}.
\]

(EC.17)

Next, we observe that \( \frac{\theta_1}{\theta_2} \leq \theta_2 \) for every feasible \( \lambda \) for Line (EC.17). Using this inequality, we obtain the following upper bound on Line (EC.17).

\[
\sup_{Q \in W_1(\xi; \theta_1, z)} E_Q [f(\xi)] \leq \sup_{0 \leq \lambda \leq 1 - \frac{\theta_1}{\theta_2}} \left\{ \lambda \left( \sup_{Q \in W_\infty(\xi; \frac{\theta_1}{2}, z)} E_Q [f(\xi)] \right) + (1 - \lambda) \left( \sup_{Q \in W_\infty(\xi; \theta_2, z)} E_Q [f(\xi)] \right) \right\}
= \sup_{Q \in W_\infty(\xi; \theta_2, z)} E_Q [f(\xi)] \sup_{0 \leq \lambda \leq 1 - \frac{\theta_1}{\theta_2}} \left\{ \lambda \left( \sup_{Q \in W_\infty(\xi; \frac{\theta_1}{2}, z)} E_Q [f(\xi)] \right) - \sup_{Q \in W_\infty(\xi; \theta_2, z)} E_Q [f(\xi)] \right\},
\]

(EC.18)

where the above equality comes from rearranging terms. For every \( \lambda \geq \frac{\theta_1}{\theta_2} \), it immediately follows from \( \frac{\theta_1}{\lambda} \leq \theta_2 \) that

\[
\sup_{Q \in W_\infty(\xi; \frac{\theta_1}{2}, z)} E_Q [f(\xi)] - \sup_{Q \in W_\infty(\xi; \theta_2, z)} E_Q [f(\xi)] \leq 0.
\]

and the above holds at equality when \( \lambda = \frac{\theta_1}{\theta_2} \). Therefore,

\[
\sup_{0 \leq \lambda \leq 1 - \frac{\theta_1}{\theta_2}} \left\{ \lambda \left( \sup_{Q \in W_\infty(\xi; \frac{\theta_1}{2}, z)} E_Q [f(\xi)] \right) - \sup_{Q \in W_\infty(\xi; \theta_2, z)} E_Q [f(\xi)] \right\}
= \sup_{0 \leq \lambda \leq \frac{\theta_1}{\theta_2}} \left\{ \lambda \left( \sup_{Q \in W_\infty(\xi; \frac{\theta_1}{2}, z)} E_Q [f(\xi)] \right) - \sup_{Q \in W_\infty(\xi; \theta_2, z)} E_Q [f(\xi)] \right\}
\leq \sup_{0 \leq \lambda \leq \frac{\theta_1}{\theta_2}} \left\{ \lambda \sup_{\zeta \in Z} [f(\zeta)] \right\}
\leq \frac{2\theta_1}{\theta_2} \sup_{\zeta \in Z} |f(\zeta)|.
\]

(EC.19)

Line (EC.19) follows because we can without loss of generality restrict \( \lambda \) to the interval \([0, \frac{\theta_1}{\theta_2}] \). Line (EC.20) is obtained by applying the global lower and upper bounds on \( f(\zeta) \). Finally, we obtain Line (EC.21) since

\[
0 \leq \sup_{\zeta \in Z} f(\zeta) - \inf_{\zeta \in Z} f(\zeta) \leq 2 \sup_{\zeta \in Z} |f(\zeta)|.
\]

Combining Lines (EC.18) and (EC.21), we obtain the desired result. □
We now present our main result on the gap between DRO using the 1- and \( \infty \)-Wasserstein ambiguity sets.

**Theorem EC.1.** Let \( Z \subseteq \mathbb{R}^d \), \( f : Z \to \mathbb{R} \) be measurable, and \( \hat{\xi}^1, \ldots, \hat{\xi}^N \in Z \). If \( \theta_2 \geq 2\theta_1 \geq 0 \), then

\[
\sup_{Q \in W_1(\hat{\xi}^1, \ldots, \hat{\xi}^N; \theta_1, Z)} \mathbb{E}_Q[f(\xi)] \leq \sup_{Q \in W_\infty(\hat{\xi}^1, \ldots, \hat{\xi}^N; \theta_2, Z)} \mathbb{E}_Q[f(\xi)] + \frac{4\theta_1}{\theta_2} \sup_{\xi \in Z} |f(\xi)|.
\]

**Proof.** We recall from (Esfahani and Kuhn 2018, Gao and Kleywegt 2016) that the 1-Wasserstein ambiguity set can be written as

\[
W_1(\hat{\xi}^1, \ldots, \hat{\xi}^N; \theta, Z) = \left\{ \frac{1}{N} \sum_{j=1}^N Q_j : \frac{1}{N} \sum_{j=1}^N \mathbb{E}_{Q_j} \left[ \| \xi - \hat{\xi}^j \| \right] \leq \theta \right\}. \tag{EC.22}
\]

Therefore,

\[
\sup_{Q \in W_1(\hat{\xi}^1, \ldots, \hat{\xi}^N; \theta_1, Z)} \mathbb{E}_Q[f(\xi)] = \sup_{\gamma \in \mathbb{R}_+^N} \left\{ \frac{1}{N} \sum_{j=1}^N \sup_{Q \in W_1(\hat{\xi}^j; \gamma, Z)} \mathbb{E}_Q[f(\xi)] : \frac{1}{N} \sum_{j=1}^N \gamma_j \leq \theta_1 \right\}. \tag{EC.23}
\]

For any choice of \( \gamma \in \mathbb{R}_+^N \), we can partition the components \( \gamma_j \) into those that satisfy \( 2\gamma_j \leq \theta_2 \) and \( 2\gamma_j > \theta_2 \). Thus,

\[
\sup_{Q \in W_1(\hat{\xi}^1, \ldots, \hat{\xi}^N; \theta_1, Z)} \mathbb{E}_Q[f(\xi)] = \sup_{\gamma \in \mathbb{R}_+^N} \left\{ \frac{1}{N} \sum_{j \in [N] : 2\gamma_j \leq \theta_2} \sup_{Q \in W_1(\hat{\xi}^j; \gamma, Z)} \mathbb{E}_Q[f(\xi)] + \frac{1}{N} \sum_{j \in [N] : 2\gamma_j > \theta_2} \sup_{Q \in W_1(\hat{\xi}^j; \gamma, Z)} \mathbb{E}_Q[f(\xi)] : \frac{1}{N} \sum_{j=1}^N \gamma_j \leq \theta_1 \right\} \tag{EC.24}
\]

where the upper bound on Line (EC.24) follows from upper bounding each DRO problem for which \( 2\gamma_j > \theta_2 \) by \( \sup_{\xi \in Z} |f(\xi)| \). Due to the constraints on \( \gamma \), there can be at most \( \frac{2N\theta_1}{\theta_2} \) components which satisfy \( 2\gamma_j > \theta_2 \). Thus, it follows from Line (EC.24) that

\[
\sup_{Q \in W_1(\hat{\xi}^1, \ldots, \hat{\xi}^N; \theta_1, Z)} \mathbb{E}_Q[f(\xi)] \leq \sup_{\gamma \in \mathbb{R}_+^N} \left\{ \frac{1}{N} \sum_{j \in [N] : 2\gamma_j \leq \theta_2} \sup_{Q \in W_1(\hat{\xi}^j; \gamma, Z)} \mathbb{E}_Q[f(\xi)] : \frac{1}{N} \sum_{j=1}^N \gamma_j \leq \theta_1 \right\} + \frac{2\theta_1}{\theta_2} \sup_{\xi \in Z} |f(\xi)|. \tag{EC.25}
\]

To conclude the proof, we apply Lemma EC.1 to each DRO problem in Line (EC.25) and obtain the following upper bounds.

\[
\sup_{Q \in W_1(\hat{\xi}^1, \ldots, \hat{\xi}^N; \theta_1, Z)} \mathbb{E}_Q[f(\xi)]
\]
Line (EC.26) follows from applying Lemma EC.1 to Line (EC.25). Line (EC.27) follows because
\[ \sup_{Q \in W_{\infty}(\xi; \theta_2, \varepsilon)} \mathbb{E}_Q[f(\xi)] + \frac{2\gamma_j}{\theta_2} \sup_{\xi \in \mathbb{Z}} |f(\xi)| \geq 0 \]
for each component that satisfies \( 2\gamma_j > \theta_2 \), and thus adding these quantities to Line (EC.26) results
in an upper bound. Finally, Line (EC.28) follows from Lemma 1 and the constraint \( \frac{1}{N} \sum_{j=1}^{N} \gamma_j \leq \theta_1 \).
This concludes the proof. \( \square \)

**EC.4.2. Proof of Theorem 4 and a generalization**

We prove a slight generalization of Theorem 4, in which \( c(\xi)^{\top} x(\xi) \) is replaced by any measurable
function \( f(\xi) \).

**Theorem 4.** Suppose Assumptions 1 and 2 hold. Then, for every \( \hat{\kappa} > 0 \), there exists a \( \bar{N} \in \mathbb{N} \),
\( \mathbb{P}^\infty \)-almost surely, such that
\[
\mathbb{E}_\mathbb{P} [ f(\xi) \mathbb{1}(\xi \in S_N) ] \leq \frac{1}{N} \sum_{j=1}^{N} \sup_{\xi \in S_N} f(\xi) + M_N \sup_{\xi \in S_N} |f(\xi)|,
\]
for all \( N \geq \bar{N} \) and all measurable functions \( f : \Xi \to \mathbb{R} \), where \( M_N \equiv N^{-\left( \frac{1}{d+1} \right) (d+2)} \log N \to 0 \) as
\( N \to \infty \).

**Proof.** Let \( \kappa > 0 \) be the coefficient from Assumption 2, and define \( \hat{\kappa} = \hat{\kappa} \kappa / 8 \). For each \( N \in \mathbb{N} \),
define
\[
\delta_N \equiv \begin{cases} \hat{\kappa} N^{-\frac{1}{d}} \log N, & \text{if } d = 1, \\ \hat{\kappa} N^{-\frac{1}{d}} (\log N)^2, & \text{if } d \geq 2. \end{cases}
\]
Then, it follows from Fournier and Guillin (2015) and Assumption 1 that \( \mathbb{P} \in W_1(\bar{\xi}^1, \ldots, \bar{\xi}^N; \bar{\delta}_N, \bar{\Xi}) \)
for all sufficiently large \( N \), \( \mathbb{P}^{\infty} \)-almost surely (see Lemma EC.2 in Section EC.4.3). Therefore, for
every measurable function \( f : \Xi \to \bar{\mathbb{R}} \),
\[
\mathbb{E}_\mathbb{P} [ f(\xi) \mathbb{1}(\xi \in S_N) ]
\]
where the inequality holds \( P^\infty \)-almost surely for all sufficiently large \( N \). Next, we observe that \( g(\xi) = 0 \) when \( \xi \notin S_N \) and is nonnegative otherwise. Therefore, without loss of generality, we can restrict the supremum over the expectation of \( g(\xi) \) to distributions with support contained in \( S_N \) (see Lemma EC.3 in Section EC.4.3), and thus Line (EC.29) is equal to

\[
\begin{align*}
\sup_{Q \in \mathcal{W}_1(\xi^1, \ldots, \xi^N; \delta_N, \Xi)} \mathbb{E}_Q \left[ \left( f(\xi) + \sup_{\zeta \in S_N} |f(\zeta)| \right) \mathbb{I}\{\xi \in S_N\} \right] - \left( \sup_{\zeta \in S_N} |f(\zeta)| \right) P(\xi \in S_N),
\end{align*}
\]

(E.C.29)

where the first equality follows because the support of the ambiguity set is restricted to those which place mass on \( S_N \), and the second equality follows because \( \sup_{\zeta \in S_N} |f(\zeta)| \) is independent of \( Q \). By Assumption 2 and the construction of \( \delta_N \), we have that \( \epsilon_N \geq 2\delta_N \) for all sufficiently large \( N \). Thus, it follows from Theorem EC.1 (see Section EC.4.1) that Line (EC.30) is upper bounded by

\[
\begin{align*}
\sup_{Q \in \mathcal{W}_\infty(\xi^1, \ldots, \xi^N, \epsilon_N; S_N)} \mathbb{E}_Q \left[ f(\xi) \right] + \frac{4\delta_N}{\epsilon_N} \sup_{\zeta \in S_N} |f(\zeta)| + \left( \sup_{\zeta \in S_N} |f(\zeta)| \right) P(\xi \notin S_N).
\end{align*}
\]

(E.C.30)

By Lemma 1, the first supremum can be replaced by \( \frac{1}{N} \sum_{j=1}^N \sup_{\zeta \in U_N} f(\zeta) \). By the definition of \( \delta_N \), and since \( \epsilon_N = \kappa N^{-\frac{1}{d}} \) when \( d = 1 \) and \( \epsilon_N = \kappa N^{-\frac{2}{d+1}} \) when \( d \geq 2 \), we have that \( \frac{4\delta_N}{\epsilon_N} \leq \frac{\kappa}{2} M_N \) for all sufficiently large \( N \). Finally, Theorem 3 (see Section 5.2 and Section EC.3) implies that \( P(\xi \notin S_N) \leq \frac{\kappa}{2} M_N \) for all sufficiently large \( N \), \( P^\infty \)-almost surely. This concludes the proof. \( \Box \)

**EC.4.3. Miscellaneous Results**

In this section, we present tedious intermediary results which are necessary for the proof of Theorem 4.

**Lemma EC.2.** Suppose Assumption 1 holds, and let

\[
\delta_N = \begin{cases}
\overline{\kappa} N^{-\frac{1}{d}} \log N, & \text{if } d = 1, \\
\overline{\kappa} N^{-\frac{1}{d}} (\log N)^2, & \text{if } d \geq 2,
\end{cases}
\]

for any fixed \( \overline{\kappa} > 0 \). Then, \( P \in \mathcal{W}_1(\xi^1, \ldots, \xi^N; \delta_N, \Xi) \) for all sufficiently large \( N \), \( P^\infty \)-almost surely.
Proof. Let $\bar{N} \in \mathbb{N}$ be any index such that $\delta_N \leq 1$ for all $N \geq \bar{N}$. It follows from Assumption 1 that there exists an $a > 1$ such that $b \equiv \mathbb{E} [ \exp ( ||\xi||^a ) ] < \infty$. Thus, it follows from (Fournier and Guillin 2015, Theorem 2) that there exist constants $c_1, c_2 > 0$ (which depend only $a$, $b$, and $d$) such that for all $N \geq \bar{N}$,

$$
\mathbb{P}^N \left( \mathbb{P} \not\in W_1 (\hat{\xi}^1, \ldots, \hat{\xi}^N; \delta_N, \Xi) \right) \leq \begin{cases} 
  c_1 \exp \left( -c_2 N \delta_N^2 \right), & \text{if } d = 1, \\
  c_1 \exp \left( -\frac{c_2 N \delta_N^2}{\log(2+1/\delta_N)^2} \right), & \text{if } d = 2, \\
  c_1 \exp \left( -c_2 N \delta_N^d \right), & \text{if } d \geq 3.
\end{cases}
$$

(EC.31)

First, suppose $d = 1$ and $N \geq \bar{N}$. Then, it follows from the definition of $\delta_N = \bar{\kappa} N^{-\frac{1}{2}} \log N$ and Line (EC.31) that

$$
\mathbb{P}^N \left( \mathbb{P} \not\in W_1 (\hat{\xi}^1, \ldots, \hat{\xi}^N; \delta_N, \Xi) \right) \leq c_1 \exp \left( -c_2 N \delta_N^2 \right) = c_1 \exp \left( -c_2 \bar{\kappa}^2 (\log N)^2 \right).
$$

Second, suppose $d = 2$ and $N \geq \bar{N}$. Then, it follows from the definition of $\delta_N = \bar{\kappa} N^{-\frac{1}{2}} (\log N)^2$ and Line (EC.31) that there exists some constant $\bar{c} > 0$ (which depends only on $\bar{\kappa}$ and $c_2$) such that

$$
\mathbb{P}^N \left( \mathbb{P} \not\in W_1 (\hat{\xi}^1, \ldots, \hat{\xi}^N; \delta_N, \Xi) \right) \leq c_1 \exp \left( -\frac{c_2 N \delta_N^2}{\log(2+1/\delta_N)^2} \right)
= c_1 \exp \left( -\frac{c_2 \bar{\kappa}^2 (\log N)^4}{\log(2+\bar{\kappa}^{-2} N^2 (\log N)^{-2})^2} \right)
\leq c_1 \exp \left( -\frac{c_2 \bar{\kappa}^2 (\log N)^4}{\log(2+\bar{\kappa}^{-2} N^2)^2} \right)
\leq c_1 \exp \left( -\bar{c} (\log N)^2 \right).
$$

Third, suppose $d \geq 3$ and $N \geq \bar{N}$. Then, it follows from the definition of $\delta_N = \bar{\kappa} N^{-\frac{1}{2}} (\log N)^2$ and Line (EC.31) that

$$
\mathbb{P}^N \left( \mathbb{P} \not\in W_1 (\hat{\xi}^1, \ldots, \hat{\xi}^N; \delta_N, \Xi) \right) \leq c_1 \exp \left( -c_2 N \delta_N^d \right) = c_1 \exp \left( -c_2 (\log N)^{2d} \right).
$$

Therefore, for any $d \geq 1$, we have shown that

$$
\sum_{N=1}^{\infty} \mathbb{P}^N \left( \mathbb{P} \not\in W_1 (\hat{\xi}^1, \ldots, \hat{\xi}^N; \delta_N, \Xi) \right) < \infty,
$$

and thus the desired result follows from the Borel-Cantelli lemma. \qed

Lemma EC.3. Suppose $\Xi \subseteq \mathbb{R}^d$ and $\hat{\xi}^1, \ldots, \hat{\xi}^N \in \mathcal{Z} \subseteq \Xi$. Let $g : \Xi \rightarrow \bar{\mathbb{R}}$ be any measurable function where $g(\xi) \geq 0$ for all $\xi \in \mathcal{Z}$. Then, for all $\theta \geq 0$,

$$
\sup_{Q \in \mathcal{W}_1 (\hat{\xi}^1, \ldots, \hat{\xi}^N; \theta, \Xi)} \mathbb{E}_Q [ g(\xi) \mathbb{I} \{ \xi \in \mathcal{Z} \} ] = \sup_{Q \in \mathcal{W}_1 (\hat{\xi}^1, \ldots, \hat{\xi}^N; \theta, \mathcal{Z})} \mathbb{E}_Q [ g(\xi) ].
$$
Proof. For compactness, define \( \bar{g}(\zeta) \equiv g(\xi) \mathbb{1} \{ \xi \in \mathcal{Z} \} \) for all \( \zeta \in \Xi \). It readily follows from \( \mathcal{Z} \subseteq \Xi \) that

\[
\sup_{\mathcal{Q} \in \mathcal{W}_1(\xi^1, \ldots, \xi^N, \theta, \Xi)} \mathbb{E}_Q [g(\xi)] \geq \sup_{\mathcal{Q} \in \mathcal{W}_1(\xi^1, \ldots, \xi^N, \theta, Z)} \mathbb{E}_Q [g(\xi)] = \sup_{\mathcal{Q} \in \mathcal{W}_1(\xi^1, \ldots, \xi^N, \theta, Z)} \mathbb{E}_Q [\bar{g}(\xi)],
\]

where the equality holds since \( \bar{g}(\zeta) = g(\zeta) \) for all \( \zeta \in \mathcal{Z} \).

It remains to show the other direction. By the Richter-Rogonsinski Theorem (see Theorem 7.32 and Proposition 6.40 of Shapiro et al. (2009)),

\[
\sup_{\mathcal{Q} \in \mathcal{W}_1(\xi^1, \ldots, \xi^N, \theta, \Xi)} \mathbb{E}_Q [\bar{g}(\xi)]
\]

\[
= \sup_{\zeta^1, \zeta^2 \in \Xi, \lambda^i \in [0,1]} \frac{1}{N} \sum_{j=1}^{N} \left( \lambda^1 \bar{g}(\zeta^{1j}) + (1 - \lambda^1) \bar{g}(\zeta^{2j}) \right)
\]

subject to

\[
\frac{1}{N} \sum_{j=1}^{N} \left( \lambda^i \| \zeta^1 - \hat{\zeta}^i \| + (1 - \lambda^i) \| \zeta^2 - \hat{\zeta}^i \| \right) \leq \theta
\]

For any arbitrary \( \eta > 0 \), let \((\hat{\zeta}^{1j}, \hat{\zeta}^{2j}, \hat{\lambda}^j)_{j \in [N]}\) be an \( \eta \)-optimal solution to the above optimization problem. We now perform a transformation on this solution. For each \( j \in [N] \), define \( \hat{\lambda}^j = \hat{\lambda}^j \), and for each \( * \in \{1,2\} \), define \( \hat{\zeta}^{*j} = \hat{\zeta}^{*j} \) if \( \hat{\zeta}^{*j} \in \mathcal{Z} \) and \( \hat{\zeta}^{*j} = \hat{\zeta}^{j} \) otherwise. Since \( \bar{g}(\zeta) \geq 0 \) for all \( \zeta \in \Xi \) and \( g(\zeta) = 0 \) for all \( \zeta \in \mathcal{Z} \), we have that \( \bar{g}(\hat{\zeta}^{*j}) \geq \bar{g}(\hat{\zeta}^{*j}) \). By construction, \((\hat{\zeta}^{1j}, \hat{\zeta}^{2j}, \hat{\lambda}^j)_{j \in [N]}\) is a feasible solution to the above optimization problem, and is also feasible for

\[
\sup_{\zeta^1, \zeta^2 \in \mathcal{Z}, \lambda^i \in [0,1]} \frac{1}{N} \sum_{j=1}^{N} \left( \lambda^i \bar{g}(\zeta^{1j}) + (1 - \lambda^i) \bar{g}(\zeta^{2j}) \right)
\]

subject to

\[
\frac{1}{N} \sum_{j=1}^{N} \left( \lambda^i \| \zeta^1 - \hat{\zeta}^i \| + (1 - \lambda^i) \| \zeta^2 - \hat{\zeta}^i \| \right) \leq \theta
\]

We have thus shown that

\[
\sup_{\mathcal{Q} \in \mathcal{W}_1(\xi^1, \ldots, \xi^N, \theta, \Xi)} \mathbb{E}_Q [\bar{g}(\xi)] \leq \frac{1}{N} \sum_{j=1}^{N} \left( \hat{\lambda}^j \bar{g}(\hat{\zeta}^{1j}) + (1 - \hat{\lambda}^j) \bar{g}(\hat{\zeta}^{2j}) \right) + \eta
\]

\[
\leq \frac{1}{N} \sum_{j=1}^{N} \left( \hat{\lambda}^j \bar{g}(\hat{\zeta}^{1j}) + (1 - \hat{\lambda}^j) \bar{g}(\hat{\zeta}^{2j}) \right) + \eta \leq \mathbb{E}_Q [\bar{g}(\xi)] + \eta.
\]

Since \( \eta > 0 \) was chosen arbitrarily, we have shown the other direction. \( \square \)

**EC.5. Proofs of Theorem 5 from Section 5.3**

For clarity of exposition, we split Theorem 5 into two parts.

**Theorem 5A.** Suppose Assumptions 1 and 2 hold. Then, \( \mathbb{P}^\infty \)-almost surely we have

\[ J \leq \liminf_{N \to \infty} \hat{J}_N. \]
Proof. For any arbitrary $\rho > 0$, it follows from Theorem 3 that $\mathbb{P}(\xi \in S_N) \geq 1 - \rho$ for all $N \in \mathbb{N}$ sufficiently large, $\mathbb{P}^\infty$-almost surely, where $S_N$ is shorthand for $\bigcup_{j=1}^N \mathcal{U}_j$. Thus, $\mathbb{P}^\infty$-almost surely,

$$J \leq \liminf_{N \to \infty} \minimize_{\epsilon \in \mathbb{F}^{\text{(EXP)}}} \mathbb{E}_P [c(\epsilon) \tr x(\epsilon)]$$

subject to $A(\zeta) x(\zeta) \leq b(\zeta), \quad (\text{EC.32})$

$\forall \zeta \in S_N.$

Next, for any feasible decision rule to (EC.32), it follows from the law of iterated expectation that

$$\mathbb{E}_P [c(\epsilon) \tr x(\epsilon)] = \mathbb{E}_P [c(\epsilon) \tr x(\epsilon) \mathbb{I} \{ \epsilon \in S_N \}] + \mathbb{E}_P [c(\epsilon) \tr x(\epsilon) \mathbb{I} \{ \epsilon \notin S_N \}].$$

We now bound these two quantities. First, for all decision rules which satisfy Condition (EXP),

$$\mathbb{E}_P [c(\epsilon) \tr x(\epsilon) \mathbb{I} \{ \epsilon \notin S_N \}] \leq \mathbb{E} [L \exp (\|\epsilon\|) \mathbb{I} \{ \epsilon \notin S_N \}] \leq \sqrt{\mathbb{E} [L^2 \exp (2\|\epsilon\|)]} \mathbb{P}(\epsilon \notin S_N),$$

where the first inequality follows from Condition (EXP) and the second inequality follows from the Cauchy-Schwartz inequality. Since Assumption 1 implies that $b \equiv \mathbb{E} [\exp (\|\epsilon\|^a)] < \infty$ for some $a > 1$, we have $\mathbb{E} [\exp (2\|\epsilon\|)] < \infty$. Thus, it follows from Theorem 3 that $\alpha_N \to 0$ as $N \to \infty$, $\mathbb{P}^\infty$-almost surely. We have therefore shown that

$$\lim_{N \to \infty} \mathbb{E}_P [c(\epsilon) \tr x(\epsilon) \mathbb{I} \{ \epsilon \notin S_N \}] = 0, \quad \mathbb{P}^\infty$$-almost surely,

where the convergence holds uniformly over all decision rules which satisfy Condition (EXP).

To bound the second term, we invoke Theorem 4. Define $M_N \equiv N^{-\frac{1}{(d+1)(d+2)}} \log N$. Then, Theorem 4 implies there exists a $\bar{N} \in \mathbb{N}$, $\mathbb{P}^\infty$-almost surely, such that

$$\mathbb{E}_P [c(\epsilon) \tr x(\epsilon) \mathbb{I} \{ \epsilon \in S_N \}] \leq \frac{1}{N} \sum_{j=1}^N \sup_{\zeta \in \mathcal{U}_j} c(\zeta) \tr x(\zeta) + M_N \sup_{\zeta \in S_N} |c(\zeta) \tr x(\zeta)|$$

for all $N \geq \bar{N}$ and decision rules. Since we are only considering decision rules which satisfy Condition (EXP), and since Assumption 1 holds, we shall now show $\beta_N$ converges to zero almost surely. Choose any arbitrary $\eta > 0$. Then, for each $N \geq \bar{N}$ and for every decision rule which satisfies Condition (EXP),

$$\mathbb{P}^N \left( M_N \sup_{\zeta \in S_N} |c(\zeta) \tr x(\zeta)| > \eta \right) \leq \mathbb{P}^N \left( M_N \sup_{\zeta \in S_N} \exp (\|\zeta\|) > \eta \right)$$

$$\leq N \mathbb{P} \left( M_N \sup_{\zeta \in \mathbb{E} \|\zeta\| \leq N} \exp (\|\zeta\|) > \eta \right)$$

$$\leq N \mathbb{P} \left( M_N \exp (\|\zeta\| + \epsilon_N) > \eta \right)$$

$$\leq N \mathbb{P} \left( \exp (\|\zeta\|^a) > \exp \left( \log \left( \frac{\eta}{M_N \exp (\epsilon_N)} \right)^a \right) \right)$$

$$\leq \exp \left( \log \left( \frac{\eta}{M_N \exp (\epsilon_N)} \right)^a \right).$$
The first inequality follows from Condition (EXP), and the second inequality follows from the union bound. The third inequality follows because, for each $\zeta \in \Xi$ that satisfies $\|\zeta - \xi\| \leq \epsilon_N$, the triangle inequality implies that $\|\zeta\| \leq \|\xi\| + \epsilon_N$. The fourth inequality follows from applying monotonic transformations to both sides of the inequality. The final inequality follows from Markov’s inequality, where $b \equiv E[\exp(\|\xi\|^a)] < \infty$ for some $a > 1$. Under Assumption 2 and the definition of $M_N$, we observe that the denominator in the final expression exceeds $N^3$ for all $N$ sufficiently large. Thus, we have shown that

$$\sum_{N=1}^{\infty} \mathbb{P}^N \left( M_N \sup_{\zeta \in S_N} |c(\zeta)^T x(\zeta)| > \eta \right) < \infty$$

for all arbitrary $\eta > 0$, and so the Borel-Cantelli lemma implies that $\beta_N \to 0$ as $N \to \infty$, $\mathbb{P}^\infty$-almost surely.

Combining the above analysis with Problem (EC.32), we obtain the desired result. □

**Theorem 5b.** Suppose Assumptions 1 and 2 hold. Then, $\mathbb{P}^\infty$-almost surely we have

$$\limsup_{N \to \infty} \hat{J}_N \leq \bar{J}.$$ 

**Proof.** Choose any arbitrary $\eta > 0$. Then, there exists a $\rho > 0$ such that

$$\bar{J} \geq \eta + \min_{\mathbf{x} \in X : x \in \text{(EXP)}} \mathbb{E}_p \left[ c(\xi)^T x(\xi) \right]$$

subject to $A(\xi) x(\xi) \leq b(\xi)$,

$$\forall \xi \in \Xi$$ such that $\|\xi - \xi\| \leq \rho$ for some $\xi' \in S$.

Let $\bar{x}(\cdot)$ be any decision rule that is feasible for the above optimization and satisfies

$$\bar{J} \geq 2\eta + \mathbb{E}_p \left[ c(\xi)^T \bar{x}(\xi) \right]. \quad \text{(EC.33)}$$

Under Assumption 2, $\epsilon_N \to 0$ as $N \to \infty$. Therefore, the aforementioned decision rule is feasible for Problem (1) for all $N$ sufficiently large. Thus,

$$\limsup_{N \to \infty} \hat{J}_N \leq \limsup_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} \sup_{\zeta \in \Xi : \|\zeta - \xi\| \leq \epsilon_N} c(\zeta)^T \bar{x}(\zeta)$$

$$\leq \lim_{k \to \infty} \limsup_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} \sup_{\zeta \in \Xi : \|\zeta - \xi\| \leq \epsilon_k} c(\zeta)^T \bar{x}(\zeta)$$

$$= \lim_{k \to \infty} \mathbb{E}_p \left[ \sup_{\zeta \in \Xi : \|\zeta - \xi\| \leq \epsilon_k} c(\zeta)^T \bar{x}(\zeta) \right] \quad \mathbb{P}^\infty-\text{almost surely}$$

$$= \mathbb{E}_p \left[ \lim_{k \to \infty} \sup_{\zeta \in \Xi : \|\zeta - \xi\| \leq \epsilon_k} c(\zeta)^T \bar{x}(\zeta) \right] \quad \text{(EC.34)}$$

$$= \mathbb{E}_p \left[ c(\xi)^T \bar{x}(\xi) \right] \quad \text{(EC.35)}$$

$$\leq \bar{J} - 2\eta. \quad \text{(EC.36)}$$
The first inequality follows because $\bar{x}(\cdot)$ is a feasible but possibly suboptimal decision rule for the optimization problem corresponding to $\hat{J}_N$. The second inequality follows because $\epsilon_k \to 0$ monotonically as $k \to \infty$. The first equality follows from the strong law of large numbers (Williams 1991, Theorem 12.10), which applies here since, for all $k \in \mathbb{N}$,

$$\mathbb{E}_p \left[ \sup_{\zeta \in \Xi, \|\zeta - \xi\| \leq \epsilon_k} c(\zeta)^T \bar{x}(\zeta) \right] \leq \mathbb{E}_p \left[ \sup_{\zeta \in \Xi, \|\zeta - \xi\| \leq \epsilon_k} |c(\zeta)^T \bar{x}(\zeta)| \right] \leq \mathbb{E}_p \left[ \exp(\|\xi\| + \epsilon_k) \right] < \infty.$$ 

Line (EC.34) follows the dominated convergence theorem (Williams 1991, Theorem 5.9), which applies here since $\epsilon_k \to 0$ as $k \to \infty$, and the right side of Line (EC.37) is integrable. Line (EC.35) follows from the definition of the upper semicontinuous envelope $\bar{E}_p[\cdot]$. Line (EC.36) follows from Line (EC.33). Since $\eta > 0$ was chosen arbitrarily, we have proven the desired result. □

**EC.6. Technical Details of Example 3 from Section 5.3**

Recall that $\mathbb{P}(\xi_1 > \alpha) = (1 - \alpha)^k$. Thus, for any $k > 0$,

$$J = \lim_{\rho \downarrow 0} \min_{x_1 \in \mathbb{Z}} \{x_1 : \mathbb{P}(x_1 \geq \xi_1) \geq 1 - \rho\} = 1, \text{ and}$$

$$\bar{J} = \lim_{\rho \downarrow 0} \min_{x_1 \in \mathbb{Z}} \{x_1 : x_1 \geq 1 + \rho\} = 2.$$

Furthermore, given the historical data $\hat{\xi}_1^1, \ldots, \hat{\xi}_1^N$, the choice of the robustness parameter $\epsilon_N = N^{-\frac{1}{2}}$, and $\Xi \equiv [0, 2]$,

$$\hat{J}_N = \min_{x_1 \in \mathbb{Z}} \{x_1 : x_1 \geq \xi_1, \forall \xi_1 \in S_N\} = \begin{cases} 1, & \text{if } \max_{j \in [N]} \hat{\xi}_1^j \leq 1 - N^{-\frac{1}{2}}, \\ 2, & \text{if } \max_{j \in [N]} \hat{\xi}_1^j > 1 - N^{-\frac{1}{2}}. \end{cases}$$

We first show that

$$\mathbb{P}^\infty \left( \limsup_{N \to \infty} \hat{J}_N = 1 \right) = \begin{cases} 0, & \text{if } 0 < k \leq 3, \\ 1, & \text{if } k > 3. \end{cases} \quad \text{(Claim 1)}$$

Indeed,

$$\mathbb{P}^\infty \left( \limsup_{N \to \infty} \hat{J}_N = 1 \right)$$

$$= \mathbb{P}^\infty \left( \max_{j \in [N]} \hat{\xi}_1^j \leq 1 - N^{-\frac{1}{2}} \text{ for all sufficiently large } N \right)$$

$$= \lim_{N \to \infty} \mathbb{P}^\infty \left( \max_{j \in [N]} \hat{\xi}_1^j \leq 1 - n^{-\frac{1}{2}} \text{ for all } n \geq N \right).$$
\[
= \lim_{N \to \infty} \mathbb{P}^N \left( \max_{j \in [N]} \hat{\xi}_1^j \leq 1 - N^{-\frac{1}{2}} \text{ and } \max_{j \in [n]} \hat{\xi}_1^j \leq 1 - n^{-\frac{1}{2}} \text{ for all } n \geq N + 1 \right)
\]

\[
= \lim_{N \to \infty} \mathbb{P}^N \left( \max_{j \in [N]} \hat{\xi}_1^j \leq 1 - N^{-\frac{1}{2}} \right) \prod_{n=N+1}^{\infty} \mathbb{P} \left( \max_{j \in [n]} \hat{\xi}_1^j \leq 1 - n^{-\frac{1}{2}} \right) \max_{j \in [n-1]} \hat{\xi}_1^j \leq 1 - (n-1)^{-\frac{1}{2}} \right) \quad (EC.38)
\]

\[
= \lim_{N \to \infty} \mathbb{P}^N \left( \max_{j \in [N]} \hat{\xi}_1^j \leq 1 - N^{-\frac{1}{2}} \right) \prod_{n=N+1}^{\infty} \mathbb{P} \left( \xi_1^j \leq 1 - n^{-\frac{1}{2}} \right) \quad (EC.39)
\]

\[
= \lim_{N \to \infty} \mathbb{P}^N \left( \xi_1 \leq 1 - N^{-\frac{k}{2}} \right)^N \prod_{n=N+1}^{\infty} \left( 1 - n^{-\frac{k}{2}} \right). \quad (EC.40)
\]

Line (EC.38) follows from the law of total probability. Line (EC.39) follows because, conditional on \(\max_{j \in [n-1]} \hat{\xi}_1^j \leq 1 - (n-1)^{-\frac{1}{2}}\), we have that \(\hat{\xi}_1^j \leq 1 - n^{-\frac{1}{2}}\) for all \(j \in [n-1]\). Line (EC.40) follows from the independence of \(\hat{\xi}_1^j, j \in \mathbb{N}\). By evaluating the limit in Line (EC.41), we conclude the proof of Claim 1.

Next we show that
\[
\mathbb{P}^\infty \left( \liminf_{N \to \infty} \hat{J}_N = 1 \right) = 1 \text{ if } k \geq 3. \quad (Claim 2)
\]

Indeed,
\[
\mathbb{P}^\infty \left( \liminf_{N \to \infty} \hat{J}_N = 1 \right) = \mathbb{P}^\infty \left( \max_{j \in [N]} \hat{\xi}_1^j \leq 1 - N^{-\frac{1}{2}} \text{ for infinitely many } N \right)
\]

\[
= \lim_{N \to \infty} \mathbb{P}^\infty \left( \max_{j \in [n]} \hat{\xi}_1^j \leq 1 - n^{-\frac{1}{2}} \text{ for some } n \geq N \right)
\]

\[
\geq \lim_{N \to \infty} \mathbb{P}^N \left( \max_{j \in [N]} \hat{\xi}_1^j \leq 1 - N^{-\frac{1}{2}} \right)
\]

\[
= \lim_{N \to \infty} \mathbb{P} \left( \xi_1 \leq 1 - N^{-\frac{1}{2}} \right)^N \quad (EC.42)
\]

\[
= \lim_{N \to \infty} \left( 1 - N^{-\frac{k}{2}} \right)^N. \quad (EC.43)
\]

Line (EC.42) follows from the independence of \(\hat{\xi}_1^j, j \in \mathbb{N}\). We observe that the limit in Line (EC.43) is strictly positive when \(k \geq 3\). It follows from the Hewitt-Savage zero-one law (see, e.g., Breiman (1992), Wang and Tomkins (1992)) that the event \(\{ \max_{j \in [N]} \hat{\xi}_1^j \leq 1 - N^{-\frac{1}{2}} \text{ for infinitely many } N \} \) happens with probability zero or one. Thus, Line (EC.43) implies that the event \(\liminf_{N \to \infty} \hat{J}_N = 1 \) must occur with probability one for \(k \geq 3\).

Finally, we show that
\[
\mathbb{P}^\infty \left( \liminf_{N \to \infty} \hat{J}_N = 1 \right) = 0 \text{ if } 0 < k < 3. \quad (Claim 3)
\]
Indeed, suppose that $0 < k < 3$. Then,

$$\sum_{N=1}^{\infty} \mathbb{P}^\infty \left( \hat{J}_N = 1 \right) = \sum_{N=1}^{\infty} \mathbb{P}^N \left( \max_{j \in [N]} \hat{\xi}^j_1 \leq 1 - N^{-\frac{k}{2}} \right) = \sum_{N=1}^{\infty} \left( 1 - N^{-\frac{k}{2}} \right)^N < \infty$$

Therefore, it follows from the Borel-Cantelli lemma that

$$\mathbb{P}^\infty \left( \liminf_{N \to \infty} \hat{J}_N = 1 \right) = \mathbb{P}^\infty \left( \max_{j \in [N]} \hat{\xi}^j_1 > 1 - N^{-\frac{1}{3}} \text{ for all sufficiently large } N \right) = 0,$$

when $0 < k < 3$, which proves Claim 3.

Combining Claims 1,2,3 with the definitions of $J$ and $\hat{J}$, we have shown the desired results.

**EC.7. An Iterative Approach for Selecting Partitions**

In this section, we describe the iterative algorithm for constructing partitions for finite adaptability in Section 3. The proposed algorithm is an extension and refinement of ideas of Bertsimas and Dunning (2016) and Postek and Hertog (2016), which were designed for adaptive optimization.

The algorithm can be described succinctly as follows. At each iteration, we start with a partition $\mathcal{P} = \{P_1, \ldots, P_K\}$ of $\Xi$ and solve the resulting approximation of Problem (1) with finite adaptability. We then find the realizations $\zeta$ in each $U^j_N \cap P_k$ which are active at the optimal decision rule, and split the previous partition recursively into a refined partition $\mathcal{P}' = \{P'_1, \ldots, P'_K\}$ by separating these worst-case realizations. As established by the previous literature, this heuristic for creating partitions has theoretical justification, which we shall extend shortly to the present setting of Problem (1).

The central purpose of this section is to adapt the partitioning algorithm and theoretical justification from the previous literature to Problem (1). Indeed, since this paper considers optimizing over $N$ uncertainty sets, the methods and theory from existing literature require modification to be tractable and justified in the present setting. In Section EC.7.1, we present the notation and theoretical justification for the partitioning heuristic. In Section EC.7.2, we present the algorithm and discuss its tractability. The resulting running times and approximation quality are demonstrated empirically in Section 6.

**EC.7.1. Refined Partitions**

In this section, we discuss some theoretical concepts about partitioning which provide justification for the algorithm in the following section. For compactness, we will henceforth denote a partition of $\Xi$ by $\mathcal{P} = \{P_1, \ldots, P_K\}$. Furthermore, we use $z(\mathcal{P})$ to denote the optimal objective value of Problem (2) with partition $\mathcal{P}$:
\[
\begin{align*}
\minimize \quad & \frac{1}{N} \sum_{j=1}^{N} v_j \\
\text{subject to} \quad & c(\zeta)^T x^k \leq v_j \\
& A(\zeta) x^k \leq b(\zeta) \quad \forall \zeta \in U_N^j \cap P_k, j \in [N], k \in K_j \\
& x^k_t = x^k_{t'} \quad \forall (k, k', t) \in T(P_1, \ldots, P_K).
\end{align*}
\]

We begin with the following definition.

**Definition EC.1.** Let \(\mathcal{P} \equiv \{P_1, \ldots, P_K\}\) be a partition of \(\Xi\). Then, a collection \(\mathcal{P}' \equiv \{P'_1, \ldots, P'_{K'}\}\) is a *refined partition* of \(\mathcal{P}\) if (i) \(\mathcal{P}'\) is a partition of \(\Xi\), and (ii) for each \(k' \in [K']\), there exists \(k \in [K]\) such that \(P'_{k'} \subseteq P_k\).

Intuitively speaking, a refined partition \(\mathcal{P}'\) of \(\mathcal{P}\) is understood to be constructed by recursively splitting some sets \(P_k \in \mathcal{P}\) of \(\Xi\). For a visualization of refined partitions, see Figure EC.1. The key observation is that a refined partition will never result in a worst optimal objective value of Problem (2) compared to the original partition.

**Proposition EC.1.** If \(\mathcal{P}\) is a partition of \(\Xi\) and \(\mathcal{P}'\) is a refined partition of \(\mathcal{P}\), then \(z(\mathcal{P}') \leq z(\mathcal{P})\).

**Proof.** Suppose \(x^1, \ldots, x^K \in \mathcal{X}\) and \(v \in \mathbb{R}^N\) are optimal for Problem (2) with partition \(\mathcal{P}\). Then, one can construct a feasible (but possibly suboptimal) solution \(\tilde{x}^1, \ldots, \tilde{x}' \in \mathcal{X}\) and \(\tilde{v} \in \mathbb{R}^N\) for Problem (2) with the refined partition \(\mathcal{P}'\) of the form \(\tilde{x}' = x^k\), where \(P'_{k'} \subseteq P_k\), and \(\tilde{v} = v\). This solution results in the same objective value of \(z(\mathcal{P})\). Since this solution is possibly suboptimal for Problem (1) with the refined partition \(\mathcal{P}'\), it follows that \(z(\mathcal{P}') \leq z(\mathcal{P})\). \(\square\)

The above result establishes that a refined partition will never result in a worse optimal objective value. On the other hand, a refined partition may not necessarily result in a strictly improved optimal objective value. Thus, we desire a characterization of a refined partitions which can lead to an improvement. Towards that end, we introduce some further notation. Let \(x^1, \ldots, x^K \in \mathcal{X}\) and \(v \in \mathbb{R}^N\) be an optimal solution of Problem (2) with a partition \(\mathcal{P} = \{P_1, \ldots, P_K\}\) of \(\Xi\). For each set \(P_k \in \mathcal{P}\), we define a corresponding active set of constraints (cuts) by

\[
\mathcal{A}(P_k) = \left\{ \zeta \in P_k ; \begin{array}{l}
\exists j \in [N] \text{ and } \zeta \in U_N^j \text{ such that } c(\zeta)^T x^k = v_j \\
\text{ or } \exists i \in [m] \text{ such that } a_i(\zeta)^T x^k = b_i(\zeta) \end{array} \right\}.
\]

In other words, \(\mathcal{A}(P_k)\) is the set of realizations \(\zeta \in P_k\) for which a corresponding constraint in Problem (2) is satisfied with equality.
Remark EC.1. Henceforth, we shall assume that all decision variables are continuous; that is, \( X = \mathbb{R}^n \). The definition of active sets, and the subsequent discussion, can be modified to address the case where some decision are integral. For a detailed discussion on addressing integer decision variables in finite adaptability, we refer the reader to (Postek and Hertog 2016, Section 4.2).

We now present a result, based on active sets, which presents necessary conditions for a refined partition to strictly improve the optimal objective value of Problem (2). Unlike the previous literature, the following necessary conditions extend to finite adaptability over multiple uncertainty sets and for general refined partitions.

**Proposition EC.2.** Let \( P \equiv \{ P_1, \ldots, P_K \} \) be a partition of \( \Xi \), and let \( P' \equiv \{ P'_1, \ldots, P'_K \} \) be a refined partition of \( P \). If \( z(P') < z(P) \), then at least one of the following two conditions hold:

(a) There exists a \( k \in [K] \) such that \( A(P_k) \not\subseteq P'_k \) for all \( k' \in [K'] \).

(b) There exists \( k, l \in [K], k', l' \in [K'], \) and \( t \in [T] \) such that \( A(P_k), A(P_l) \not= \emptyset, A(P_k) \subseteq P'_k, A(P_l) \subseteq P'_l \), \( (k, l, t) \in T(P_1, \ldots, P_K) \) and \( (k', l', t) \notin T(P'_1, \ldots, P'_{K'}) \).

**Proof.** We begin with the observation that one can remove all constraints of Problem (2) that do not correspond to realizations in the active sets without impacting the optimal objective value. That is,

\[
\begin{align*}
z(P) &= \min_{v \in \mathbb{R}^N, x^k \in X} \frac{1}{N} \sum_{j=1}^{N} v_j \\
\text{subject to} & \quad c(\zeta)^T x^k \leq v_j \quad \forall \zeta \in U_j \cap A(P_k), j \in [N], k \in K_j \\
& \quad A(\zeta) x^k \leq b(\zeta) \quad \forall \zeta \in A(P_k), k \in [K] \\
& \quad x^k_t = x^l_t \quad (k, l, t) \in T(P_1, \ldots, P_K).
\end{align*}
\]  

(EC.44)

Next, consider any \( k \in [K] \) for which \( A(P_k) = \emptyset \). In this case, \( x^k \) only appears in Problem (EC.44) in constraints of the form

\[
x^k_t = x^l_t, \quad (k, l, t) \in T(P_1, \ldots, P_K).
\]  

(EC.45)

We now show that these constraints involving \( x^k \) can be removed without impacting the optimal objective value. First, we observe that the set \( T(P_1, \ldots, P_K) \) is transitive, in the sense that \( (k, l, t), (k, s, t) \in T(P_1, \ldots, P_K) \) implies \( (l, s, t) \in T(P_1, \ldots, P_K) \). Then, it follows that removing the constraints involving \( x^k \) from Problem (EC.44) does not result in eliminating non-anticipatory
In order to show that Condition (b) is indeed a necessary alternate condition, we provide an example. It thus follows from Proposition EC.1 that

$$z(\mathcal{P}) = \text{minimize}_{x \in \mathbb{R}^N, \nabla x \in \mathcal{A}} \frac{1}{N} \sum_{j=1}^{N} v_{ij}$$

subject to

$$c(\xi)^{\top} x^{k} \leq v_{ij} \quad \forall \xi \in \mathcal{U}^{j}_{N} \cap \mathcal{A}(P_{k}), \; j \in [N], \; k \in \mathcal{K}_{j}$$

$$A(\xi)x^{k} \leq b(\xi) \quad \forall \xi \in \mathcal{A}(P_{k}), \; k \in [K]$$

$$x_{t}^{i} = x_{t}^{(k)} \quad (k, l, t) \in \mathcal{T}(P_{1}, \ldots, P_{K}), \; \mathcal{A}(P_{k}), \mathcal{A}(P_{l}) \neq \emptyset.$$  (EC.46)

Let $$\mathcal{J} = \{k \in [K] : \mathcal{A}(P_{k}) \neq \emptyset\}$$. Suppose $$\mathcal{P}'$$ is a refined partition of $$\mathcal{P}$$ which satisfies neither Conditions (a) nor (b). Since Condition (a) does not hold, there exists a mapping $$\iota : \mathcal{J} \rightarrow [K']$$ such that $$\mathcal{A}(P_{k}) \subseteq P'_{\iota(k)} \subseteq P_{k}$$ for every $$k \in \mathcal{J}$$. Moreover, since Condition (b) does not hold, $$(i(k), i(l), t) \in \mathcal{T}(P'_{1}, \ldots, P'_{K})$$ whenever $$k, l \in \mathcal{J}$$ and $$P'_{K}.$$ Thus, the following is a relaxation of Problem (2) applied on partition $$\mathcal{P}'$$.

$$\tilde{z}(\mathcal{P}') = \text{minimize}_{x^{(k)} \in \mathcal{K}, \nabla x \in \mathbb{R}^N} \frac{1}{N} \sum_{j=1}^{N} v_{ij}$$

subject to

$$c(\xi)^{\top} x^{(k)} \leq v_{ij} \quad \forall \xi \in \mathcal{U}^{j}_{N} \cap \mathcal{A}(P_{k}), \; j \in [N], \; k \in \mathcal{K}_{j}$$

$$A(\xi)x^{(k)} \leq b(\xi) \quad \forall \xi \in \mathcal{A}(P_{k}), \; k \in [K],$$

$$x_{t}^{(k)} = x_{t}^{(l)} \quad (k, l, t) \in \mathcal{T}(P_{1}, \ldots, P_{K}), \; \mathcal{A}(P_{k}), \mathcal{A}(P_{l}) \neq \emptyset.$$  (EC.47)

However, Problem (EC.47) is identical to Problem (EC.46) and thus,

$$z(\mathcal{P}) = \tilde{z}(\mathcal{P}') \leq z(\mathcal{P}').$$

It thus follows from Proposition EC.1 that $$z(\mathcal{P}) = z(\mathcal{P}')$$ if Conditions (a) and (b) do not hold. □

In order to show that Condition (b) is indeed a necessary alternate condition, we provide an example.

**Example EC.1.** Consider a problem with $$T=3$$ with parameters

$$A(\xi) = \begin{bmatrix} 0 & 2 + \xi_1 & -\xi_2 \\ 0 & 0 & -1 \\ 0 & 0 & -1 \end{bmatrix}, \quad b(\xi) = \begin{bmatrix} 0 \\ -\frac{\xi_1 + \xi_2}{\xi_1 + \xi_2} \\ \frac{\xi_1 + \xi_2}{2} \end{bmatrix}, \quad c(\xi) = \begin{bmatrix} 1 & -1 - \xi_1 \end{bmatrix},$$

and $$\mathcal{X} = \mathbb{R}^3$$. Suppose there is no third-stage uncertain parameters (i.e., $$\xi = (\xi_1, \xi_2)$$), there is a single data point (i.e., $$N = 1$$), $$\Xi \equiv \mathcal{U}_1 = [-1, 1]^2$$, and we start with a $$\mathcal{P} = \{P_1, P_2\}$$ of the form

$$P_1 = \left\{ \xi : \xi_1 \in [-1, 1], \xi_2 \in \left[\frac{1}{2}, 1\right]\right\}; \quad P_2 = \left\{ \xi : \xi_1 \in [-1, 1], \xi_2 \in \left[-1, \frac{1}{2}\right]\right\}.$$
Hence, the associated non-anticipatory constraints are $\mathcal{T}(P_1, P_2) = \{(1, 2, 1), (1, 2, 2)\}$. Then, applying these parameters to Problem (2), we obtain

$$
\begin{align*}
z(P_1, P_2) &= \text{minimize} \quad v \\
\text{subject to} \quad x_2^1 &= x_2^2 \\
(1 - \zeta_1)x_2^k + 2\zeta_2x_2^k &\leq v, \\
(2 + \zeta_1)x_2^k - \zeta_2x_3^k &\leq 0, \\
\frac{\zeta_1 + \zeta_2}{2} &\leq x_3^k, \\
-\frac{\zeta_1 + \zeta_2}{2} &\leq x_3^k, \\
\forall \zeta \in P_k, \ k = 1, 2.
\end{align*}
$$

Since $P_k = P_k \cap U_1^k$ are bounded polyhedral sets, there exist binding realizations for each constraint at extreme points. Thus, we can reformulate the problem as

$$
\begin{align*}
\text{minimize} \quad v \\
\text{subject to} \quad \max\{-2x_2, 0\} + \max\{x_3^1, 2x_3^1, x_3^2, -2x_3^2\} &\leq v \\
\max\{3x_2, x_2\} + \max\{-\frac{1}{2}x_3^1, -x_3^1\} &\leq 0 \\
\max\{3x_2, x_2\} + \max\{-\frac{1}{2}x_3^2, x_3^2\} &\leq 0 \\
x_3^1, x_3^2 &\geq 1.
\end{align*}
$$

The optimal value of this problem is 4, and the optimal solution is $x_2^2 = x_2^1 = -1$ and $x_3^1 = x_3^2 = 1$. The active sets are $\mathcal{A}(P_1) = \{(1, 1)\}$, $\mathcal{A}(P_2) = \{(-1, -1)\}$. Since both of the active sets are singletons, there is no refined partition which satisfies Condition (a) in Proposition EC.2.

We now construct a refined partition that satisfies Condition (b) which improves the optimal objective value. Let $P'_1 = \{\zeta : \zeta_1 \in [0, 1], \zeta_2 \in [\frac{1}{2}, 1]\}$, $P'_2 = \{\zeta : \zeta_1 \in [-1, 0], \zeta_2 \in [\frac{1}{2}, 1]\}$, $P'_3 = \{\zeta : \zeta_1 \in [0, 1], \zeta_2 \in [-1, \frac{1}{2}]\}$, and $P'_4 = \{\zeta : \zeta_1 \in [-1, 0], \zeta_2 \in [-1, \frac{1}{2}]\}$. Note that these sets form a refined partition $\mathcal{P}'$ of the original partition. The non-anticipatory constraints for the refined partition has the form

$$
\mathcal{T}(P'_1, P'_2, P'_3, P'_4) = \{(1, 3, 2), (2, 4, 2)\} \cup \{(k, l)\}_{k,l \in [4], k \leq l}.
$$

Thus, the resulting optimization problem must enforce that $x_2^1 = x_2^3$, $x_2^1 = x_2^4$, and $x_2^k = x_2^l$ for all $k, l \in [4]$. Using evaluating the binding constraints at the extreme points, the refined partition
results in the following reformulation.

\[
\text{minimize } \quad v \\
\text{subject to } \quad \max\{-x_2^1, -2x_2^1\} + \max\{x_3^1, 2x_3^1, x_3^3, -2x_3^3\} \leq v \\
\max\{2x_2^1, 3x_2^1\} + \max\left\{-\frac{1}{2}x_3^1, -x_3^1\right\} \leq 0 \\
\max\{2x_2^1, 3x_2^1\} + \max\left\{-\frac{1}{2}x_3^3, x_3^3\right\} \leq 0 \\
\max\{-x_2^2, 0\} + \max\{2x_3^2, x_3^4\} \leq v \\
\max\{x_2^2, 2x_2^2\} + \max\left\{-\frac{1}{2}x_3^2, -x_3^2\right\} \leq 0 \\
\max\{x_2^2, 2x_2^2\} + \max\left\{-\frac{1}{2}x_3^4, x_3^4\right\} \leq 0 \\
1 \leq x_3^1, x_3^3, \frac{1}{2} \leq x_3^2, \frac{3}{4} \leq x_3^4
\]

\[(\text{EC.48})\]

An optimal solution for this problem is \(x_2^1 = x_3^3 = -\frac{3}{8}, x_2^2 = x_4^4 = -1, x_3^1 = x_4^1 = 1, x_3^2 = \frac{1}{2}, \) and \(x_3^3 = \frac{3}{4},\) with an optimal objective value of \(\frac{11}{4} < 4.\) \(\square\)

**EC.7.2. Partitioning Algorithm**

Using the results on refined partitions from the previous section, we now describe our algorithm. For simplicity, we focus on problems where \(d_t = 1\) for all \(t = 1, \ldots, T,\) although all following results can be readily extended to the general case.

We suggest an iterative partitioning algorithm which is similar to that proposed by Bertsimas and Dunning (2016). At each iteration of the algorithm, we first solve Problem (2) using a given partition \(\mathcal{P}.\) We then compute the active sets for each \(P_k \in \mathcal{P}.\) Lastly, we construct a refined partition by separating points from the active sets while satisfying the necessary conditions from Proposition EC.2. The next iteration begins with the refined partition as the \(\mathcal{P}.\)

In order to describe our heuristic for constructing the refined partition, we introduce some necessary notation. First, let us assume that the assumptions from Proposition 3 hold. Then, we can compactly represent each \(P_k \in \mathcal{P}\) by \(P_k = [\ell^k, u^k],\) where the lower bound vector \(\ell^k \in \mathbb{R}^d\) and the upper bound vector \(u^k\) may have components which equal \(-\infty\) or \(+\infty,\) respectively. Next, given a partition \(\mathcal{P},\) we define its projection onto stage \(t\) by

\[
\mathcal{P}_t = \{[\ell, u] \subseteq \mathbb{R}^t : \exists k \in [K], P_k \equiv [\ell^k, u^k], [\ell, u] = [\ell_{1:t}, u_{1:t}]\}.
\]

We now propose our heuristic for creating the refined partition, which is visualized in Figure EC.1. In the proposed heuristic, we consider each stage \(t = 1, \ldots, T\) consecutively. For each
stage $t$ and each hyperrectangle $[\hat{\ell}, \hat{u}] \in \mathcal{P}_t$, we find all of the realizations $\zeta \in \mathcal{A} = \bigcup_{k \in [K]} \mathcal{A}(P_k)$ for which $\zeta_{1:t} = (\zeta_1, \ldots, \zeta_t)$ falls inside $[\hat{\ell}, \hat{u}]$. We compute the median $M$ of the $t$th components of these realizations. Then, for each $P_k \in \mathcal{P}$ which projects to the hyperrectangle $[\hat{\ell}, \hat{u}]$, we split $P_k$ along the cut $\zeta_t = M$. The entire algorithm is presented formally in Algorithm 1.

**Algorithm 1: Iterative Data-driven Partitioning**

**Algorithm Partitioning**

**Input:** Number of iterations ITER, $\{U^j_N\}_{j \in [N]}$

**Output:** Partition $\mathcal{P} \equiv \{P_1, \ldots, P_K\}$

Initialize: $K := 1$, $\ell^1_t := \min_{\zeta \in U^1_N} \zeta_t$, $u^1_t := \max_{\zeta \in U^1_N} \zeta_t$, $P_1 := [\ell^1, u^1]$, $\mathcal{P} := \{P_1\}$

for $it = 1, \ldots, \text{ITER}$ do

Solve Problem (2) with partition $\mathcal{P} \equiv \{P_1, \ldots, P_K\}$, and generate an active set $\mathcal{A} := \bigcup_{k \in [K]} \mathcal{A}(P_k)$

for $t = 1, \ldots, T$ do

Update $(\mathcal{P}, K) := \text{AddCuts}(\mathcal{P}, \mathcal{A}, t)$

end

end

**Procedure AddCuts($\mathcal{P} \equiv \{P_1, \ldots, P_K\}, \mathcal{A}, t$)**

Initialize $\bar{K} := K$, $\bar{\mathcal{P}} := \emptyset$

for $k \in [K]$ do

Set $\bar{P}_k \equiv [\hat{\ell}^k, \hat{u}^k] := [\ell^k_{1:t}, u^k_{1:t}]$

Update $\bar{\mathcal{P}} := \bar{\mathcal{P}} \cup \bar{P}_k$

end

for $[\bar{\ell}, \bar{u}] \in \bar{\mathcal{P}}$ do

Find $\mathcal{B} := \{\zeta \in \mathcal{A} : \zeta_{1:t} \in [\bar{\ell}, \bar{u}]\}$

if $|\mathcal{B}| \geq 2$ then

Let $v$ be the ordering of the set $\{\zeta_t : \zeta_{1:t} \in \mathcal{B}\}$

Find $i$ such that $v_i$ is closest to the median of $v$ and $v_i < v_{i+1}$

Set $\text{med} := (v_i + v_{i+1})/2$

for $k \in [K]$ such that $[\ell^k_{1:t}, u^k_{1:t}] = [\bar{\ell}, \bar{u}]$ do

Set $\bar{u} := u^k$, $u^k_t := \text{med}$, $\bar{P}_k \equiv [\ell^k, \bar{u}]$

Set $K := K + 1$ and $\bar{\ell}_t := \ell^k$, $\bar{\ell}_t := \text{med}$, $\bar{P}_K \equiv [\bar{\ell}, \bar{u}]$

Update $\bar{\mathcal{P}} := \bar{\mathcal{P}} \cup P_K$

end

end

return $\bar{\mathcal{P}}, K$
While we do not suggest that Algorithm 1 is necessarily the best cut-based recursive partitioning scheme for Problem (1), it does have some desirable properties. In particular, Algorithm 1 has a benefit of decoupling the scaling of $K$ in each iteration from $N$. Additionally, in contrast to the Voronoi technique proposed by Bertsimas and Dunning (2016), our proposed mechanism forces the introduction of splits at every stage $t \in [T]$. Moreover, similarly to the algorithm in Bertsimas and Dunning (2016), the resulting refined partition has a chance of improving the optimal objective value. Indeed, Algorithm 1 ensures that, at each stage $t$, at least one of the following occurs:

(i) At least two active points in each $P_k$ will lie in two distinct sets in the refined partition, satisfying Condition (a) in Proposition EC.2.

(ii) Two active points in two sets $P_k$ and $P_l$ such that $(k, l, t) \in \mathcal{T}(P_1, \ldots, P_K)$ are separated into two new sets $P'_k$ and $P'_l$ in the refined partition such that $(k', l', t) \notin \mathcal{T}(P'_1, \ldots, P'_K)$, satisfying Condition (b) in Proposition EC.2.
Figure EC.1  Partitioning algorithm

Note. We consider an example with 6 data points, resulting in 6 balls (more specifically, hypercubes). Figures (a)-(c) depict the stages of the first iteration of the partitioning algorithm, and in Figures (d)-(f) depict the stages of the second iteration. Figure (a) shows the active points for the partition $\mathcal{P} = \{\Xi\}$, where the red points (A-E) are associated with the epigraph constraints and the green points (1-3) are associated with the regular constraints. Figure (b) shows the cut added on the first iteration at stage $t = 1$. This cut is the mid point between point A and point F on the $\zeta_1$ axis. This cut generates a refined partition of $K = 2$ sets. Figure (c) shows the cuts added in the first iteration at stage $t = 2$. For $P_1$ and $P_2$, we add a cut associated with some $\zeta_2$ separating them into two new sets. Thus, the cut on the left is the midpoint between D and E on $\zeta_2$ axis, and the cut on the right is the midpoint between B and C on $\zeta_2$ axis. Thus we end up with a refined partition with four sets. Figure (d) shows the active points for the four sets obtained in Figure (c), as well as the projection of this sets for $t = 1$. The red points (a-e) and e’ are associated with the epigraph constraints and the green points (I-V) are associated with the regular constraints. Points e and e’ generate the same worst case objective value for the same ball in the two different projected sets, and thus are both active points. Figure (e) shows the cuts added at stage $t = 1$ of the second iteration. At this stage we refine the projected sets 1 and 2. Projected set 1 contains active points a,d,e,e’,I,II, and IV, and a cut is added on the mid point between point a and point II on the $\zeta_1$ axis. Similarly, projected set 2 contains active points b,c,f,III and V, and a cut is added on the mid point between point b and point c on the $\zeta_1$ axis. These added cuts result in a refined partition containing a total of eight sets. Figure (f) shows the cuts added at stage $t = 2$ of the second iteration. Notice that sets that have fewer than two distinct points on the $\zeta_2$ axis are not partitioned further. Thus, at this stage we add a total of three horizontal cuts between points d and I, e and II, and c and III, resulting in a total of 12 sets in the final partition.
EC.8. Finite Adaptability for Problem (6) using the 1-Wasserstein Ambiguity Set

In this section, we present the extension of finite adaptability to approximately solve Problem (6) using the 1-Wasserstein ambiguity set, as suggested in Section 4.3.

We briefly review the necessary notation from Sections 3 and 4. In this section, we consider the DRO approach to multi-stage linear optimization (Problem (6) from Section 4.2), which has the form

\[
\begin{align*}
\text{minimize} & \quad \mathbb{E}_Q [c(\xi)\trans x(\xi)] \\
\text{subject to} & \quad Q(A(\xi)x(\xi) \leq b(\xi)) = 1 \\
& \quad \forall Q \in A.
\end{align*}
\]

In particular, we focus on the particular case for which the ambiguity set \(A\) of probability distributions is defined as the 1-Wasserstein ambiguity set, i.e.,

\[
A = \{ Q \in \mathcal{P}(\Xi) : d_1(Q, \hat{P}_N) \leq \epsilon_N \},
\]

where \(\hat{P}_N\) is the empirical distribution of the historical data \(\hat{\xi}^1, \ldots, \hat{\xi}^N \in \Xi \subseteq \mathbb{R}^d\), \(\epsilon_N \geq 0\) is the robustness parameter, and \(d_1 : \mathcal{P}(\Xi) \times \mathcal{P}(\Xi) \to \mathbb{R}_+ \cup \{\infty\}\) is the 1-Wasserstein distance (see Section 4). We shall assume in this section that \(\epsilon_N > 0\), in which case it follows from Theorem 2 (see Section 4.2) that the DRO problem is equivalent to

\[
\begin{align*}
\text{minimize} & \quad \mathbb{E}_Q [c(\xi)\trans x(\xi)] \\
\text{subject to} & \quad A(\zeta)x_k(\zeta) \leq b(\zeta) \\
& \quad \forall \zeta \in \Xi, k \in [K].
\end{align*}
\]

We extend the approach of finite adaptability from Section 3 to approximate Problem (EC.49). In this approach, we construct a partition \(P_1, \ldots, P_K\) of \(\Xi\), and restrict the decision rules to those which are piecewise static decision rules over the partitions. Applying the same approach to Problem (EC.49), we obtain

\[
\begin{align*}
\text{minimize} & \quad \mathbb{E}_Q \left[ \max_{k \in [K]} c(\xi)\trans x^k(\xi) \right] \\
\text{subject to} & \quad A(\zeta)x^k(\zeta) \leq b(\zeta) \\
& \quad \forall \zeta \in P_k, k \in [K] \\
x^k_t = x^{k'}_t & \quad (k, k', t) \in T(P_1, \ldots, P_K).
\end{align*}
\]
Given any $x^1, \ldots, x^K \in \mathcal{X}$, the objective function $f(\xi) = \max_{\xi \in [K]}: \xi \in P_k \ c(\xi)^\top x^k$ is upper semicontinuous in $\xi \in \Xi$. Thus, it follows from (Gao and Kleywegt 2016, Theorem 1) that Problem (EC.50) with the 1-Wasserstein ambiguity set has the following epigraph form:

$$\begin{align*}
\min_{x \in \mathcal{X}, \nu \in \mathbb{R}^N, \lambda \in \mathbb{R}^+} & \quad \lambda \epsilon_N + \frac{1}{N} \sum_{j=1}^{N} v_j \\
\text{subject to} & \quad c(\xi)^\top x^k - \lambda \|\xi - \hat{\xi}^j\| \leq v_j \quad \forall \xi \in P_k, \ j \in [N], \ k \in [K] \\
& \quad A(\xi)x^k \leq b(\xi) \quad \forall \xi \in P_k, \ k \in [K] \\
& \quad x^k = x^{k'} \quad (k, k', t) \in \mathcal{T}(P_1, \ldots, P_K).
\end{align*}$$

(EC.51)

We shall now demonstrate that Problem (EC.51) can be reformulated as a finite-dimensional optimization problem using similar techniques to Proposition 2.

**Proposition EC.3.** Let $\Xi$ and $P_1, \ldots, P_K$ be hyperrectangles of the form $\{\xi \in \mathbb{R}^d : \ell \leq \xi \leq u\}$ for $-\infty \leq \ell \leq u \leq \infty$, and let the norm $\|\cdot\|$ denote the $\ell_1$ norm. Then, Problem (EC.51) can be reformulated by adding at most $O(\text{Kmd})$ additional continuous decision variables and $O(NK + \text{Kmd})$ additional linear constraints. The reformulation is

$$\begin{align*}
\min_{\nu \in \mathbb{R}^N, x^k \in \mathcal{X}, \lambda \geq 0, \alpha^{kp}, \beta^{kp} \in \mathbb{R}^+} & \quad \lambda \epsilon_N + \frac{1}{N} \sum_{j=1}^{N} v_j \\
\text{subject to} & \quad (C^i \hat{\xi}^j)^\top x^k + \sum_{p=1}^{3} \sum_{l \in \xi^{kp}} \alpha^{kp}_i (u^k_l - \hat{\xi}^j_l) + \beta^{kp}_i (\hat{\xi}^j_l - \ell^k_l) \leq v_j - c^0 \cdot x^k \quad j \in [N], \ k \in [K] \\
& \quad \|C^i x^k - \alpha^{kp} + \beta^{kp}\|_{\infty} \leq \lambda \quad p \in [3], \ k \in [K] \\
& \quad u^k \cdot \mu^k_i - \ell^k \cdot \theta^k_i \leq b^0 - a^0 \cdot x^k \quad k \in [K], \ i \in [m] \\
& \quad \mu^k_i - \theta^k_i = (A^i)^\top x^k - b_i \quad k \in [K], \ i \in [m] \\
& \quad x^k = x^{k'} \quad (k, k', t) \in \mathcal{T}(P_1, \ldots, P_K),
\end{align*}$$

where $b_i$ is the $i$th row of matrix $B$, the fixed vectors $\ell^k, u^k$ satisfy $P_k = \{\xi \in \mathbb{R}^d : \ell^k \leq \xi \leq u^k\}$, and $I^k_1 = \{l \in [d] : \hat{\xi}^j_l < \ell^k_l\}$, $I^k_2 = \{l \in [d] : \hat{\xi}^j_l > u^k_l\}$, $I^k_3 = \{l \in [d] : \ell^k_l \leq \hat{\xi}^j_l \leq u^k_l\}$. We remark that this reformulation holds even when $\mathcal{X}$ enforces that some decisions are integral.

In order to prove Proposition EC.3, we require the following intermediary lemma.

**Lemma EC.4.** Consider the optimization problem

$$\begin{align*}
\min_{\alpha, \beta \in \mathbb{R}^+} & \quad \alpha a + \beta b \\
\text{subject to} & \quad |d - \alpha + \beta| \leq \lambda,
\end{align*}$$

(EC.52)
where \( a + b \geq 0 \) and \( \lambda \geq 0 \) are constants. Then, there exists an optimal solution \((\alpha^*, \beta^*)\) of the form
\[
(\alpha^*, \beta^*) = \begin{cases} 
([d - \lambda]_+, [\lambda - d]_+), & \text{if } a > 0, b \leq 0, \\
([d + \lambda]_+, [-d - \lambda]_+), & \text{if } a \leq 0, b > 0, \\
([d - \lambda]_+, [-d - \lambda]_+), & \text{if } a, b \geq 0.
\end{cases}
\]

Proof. When \( a + b = 0 \), any feasible solution to Problem (EC.52) is optimal and thus the desired result follows immediately.

Assume for the remainder of the proof that \( a + b > 0 \). We start by showing that any optimal solution \((\alpha^*, \beta^*)\) to Problem (EC.52) must satisfy \( \alpha^* = 0 \) or \( \beta^* = 0 \). Indeed, consider any feasible solution \((\alpha, \beta)\) to Problem (EC.52) where \( \alpha > 0 \) and \( \beta > 0 \). Define \( \epsilon = \min\{\alpha, \beta\} \) and \((\tilde{\alpha}, \tilde{\beta}) = (\alpha - \epsilon, \beta - \epsilon)\). We observe that \( \tilde{\alpha} \geq 0, \tilde{\beta} \geq 0, \) and
\[
|d - \tilde{\alpha} + \tilde{\beta}| = |d - \alpha + \beta| \leq \lambda,
\]
which implies that \((\tilde{\alpha}, \tilde{\beta})\) is feasible for Problem (EC.52). However,
\[
a\tilde{\alpha} + b\tilde{\beta} = a\alpha^* + b\beta^* - \epsilon(a + b) < a\alpha^* + b\beta^*,
\]
where the last inequality follows from \( a + b > 0 \). Thus, \((\alpha, \beta)\) cannot be an optimal solution, which implies that every optimal solution \((\alpha^*, \beta^*)\) must satisfy \( \alpha^* = 0 \) or \( \beta^* = 0 \).

We now consider different conditions for \( a \) and \( b \). First, suppose \( a > 0 \) and \( b \leq 0 \). In this case, we want \( \alpha \) to be as small as possible and \( \beta \) to be as large as possible. If \( d > \lambda \), then an optimal solution \((\alpha^*, \beta^*)\) must satisfy \( \alpha^* = d - \lambda > 0 \) in order to satisfy the constraint in Problem (EC.52). Since \( \alpha^* \) and \( \beta^* \) cannot both be positive, this implies that \( \beta^* = 0 \). Alternatively, if \( d \leq \lambda \), then \( \alpha^* = 0 \) and \( \beta^* = \lambda - d \geq 0 \) is the optimal solution by making \( \alpha^* \) as small and \( \beta^* \) as large as possible. Thus, we have shown that \( \alpha^* = [d - \lambda]_+ \) and \( \beta^* = [\lambda - d]_+ \) is the unique optimal solution. The case where \( a \leq 0 \) and \( b > 0 \) follows by identical reasoning as above. Next, suppose \( a, b \geq 0 \). In this case, we want both \( \alpha \) and \( \beta \) to be as small as possible. Therefore, if \( d > 0 \), then the optimal solution is \( \alpha^* = [d - \lambda]_+ \) and \( \beta^* = 0 \). Otherwise, the optimal solution is given by \( \alpha^* = 0 \) and \( \beta^* = [-d - \lambda]_+ \). In either case, we can write the solution compactly as \( \alpha^* = [d - \lambda]_+ \) and \( \beta^* = [-d - \lambda]_+ \). Since \( a + b > 0 \), we do not have a case where \( a \leq 0 \) and \( b \leq 0 \). Combining the different conditions for \( a \) and \( b \), we obtain the desired result. \( \square \)

Proof of Proposition EC.3. Following similar reasoning to Proposition 2, the constraints
\[
A(\zeta)x^k \leq b(\zeta) \quad \forall \zeta \in P_k, k \in [K]
\]
are satisfied if and only if there exist \( \mu_i^k, \theta_i^k \in \mathbb{R}^d_+ \) for each \( k \in [K] \) and \( i \in [m] \) which satisfy
\[
\mu_i^k - \theta_i^k = (A_i^k)\top x^k - b_i, \quad \forall k \in [K], i \in [m],
\]
where
\[
u_i^k \cdot \mu_i^k - \ell_i^k \cdot \theta_i^k \leq b_i^0 - a_i^0 \cdot x^k, \quad \forall k \in [K], i \in [m],
\]
The remainder of the proof focuses on the epigraph constraints. Recall that $c(\zeta) = c^0 + C\zeta$, $P_k$ is a hyperrectangle with bounds $\ell^k$ and $u^k$, and the norm $\| \cdot \|$ in the 1-Wasserstein distance is the $\ell_1$ norm. Then, for every $j \in [N]$ and $k \in [K]$, the corresponding epigraph constraint in Problem (EC.51) is equivalent to

$$
\sup_{\ell^k \leq \zeta \leq u^k} (C\zeta)^\top x^k - \lambda \| \zeta - \hat{\zeta}^i \|_1 \leq v_j - c^0 \cdot x^k.
$$

(EC.53)

It follows from strong duality that the left-hand side of Line (EC.53) is equivalent to

$$
\min_{\alpha, \beta \in \mathbb{R}_+^d} (C\hat{\xi}^j)^\top x^k + \alpha \cdot (u^k - \hat{\xi}^j) + \beta \cdot (\hat{\xi}^j - \ell^k)
$$

subject to $\|C^\top x^k - \alpha + \beta\|_\infty \leq \lambda$.

(EC.54)

Remark: For any index $l$ such that $u^k_l = \infty$ (alternatively, $\ell^k_l = -\infty$), the corresponding decision variable $\alpha_l$ (alternatively, $\beta_l$) should be set to zero and the term $\alpha_l(u^k_l - \hat{\xi}^j_l)$ (alternatively, $\beta_l(\hat{\xi}^j_l - \ell^k_l)$) should be dropped from the objective.

We observe that the optimal objective value of Problem (EC.54) is equal to the sum over the optimal objective values of $d$ separate optimization problems, each of the form

$$
\min_{\alpha_l, \beta_l \in \mathbb{R}_+} \alpha_l(u^k_l - \hat{\xi}^j_l) + \beta_l(\hat{\xi}^j_l - \ell^k_l)
$$

subject to $\|C^\top x^k - \alpha_l + \beta_l\| \leq \lambda$.

(EC.55)

Since $(u^k_l - \hat{\xi}^j_l) - (\ell^k_l - \hat{\xi}^j_l) \geq 0$, Lemma EC.4 implies that the optimal value of $\alpha_l$ and $\beta_l$ can be determined by checking whether $\hat{\xi}^j_l < \ell^k_l$, $\hat{\xi}^j_l > u^k_l$, or $\ell^k_l \leq \hat{\xi}^j_l \leq u^k_l$. This information is captured by the index sets

$$
\mathcal{I}^{kj}_1 = \{l \in [d] : \hat{\xi}^j_l < \ell^k_l\},
$$

$$
\mathcal{I}^{kj}_2 = \{l \in [d] : \hat{\xi}^j_l > u^k_l\},
$$

$$
\mathcal{I}^{kj}_3 = \{l \in [d] : \ell^k_l \leq \hat{\xi}^j_l \leq u^k_l\}.
$$

Thus, for each $j \in [N]$ and $k \in [K]$, Problem (EC.54) is equivalent to

$$
\min_{\alpha^p, \beta^p \in \mathbb{R}_+^d, p \in [3]} (C\hat{\xi}^j)^\top x^k + \sum_{p=1}^3 \sum_{l \in \mathcal{I}^{kj}_p} \alpha^p_l(u^k_l - \hat{\xi}^j_l) + \beta^p_l(\hat{\xi}^j_l - \ell^k_l)
$$

subject to $\|C^\top x^k - \alpha^p + \beta^p\|_\infty \leq \lambda$.

(EC.56)

The key observation is the optimal solutions $\alpha^p, \beta^p$ for this optimization problem is independent of $j \in [N]$. Thus, the constraints in Line (EC.53) are satisfied if and only if there exists $\alpha^k \in \mathbb{R}_+^d$ and $\beta^k \in \mathbb{R}_+^d$ for each $k \in [K]$ and $p \in [3]$ which satisfy

$$
(C\hat{\xi}^j)^\top x^k + \sum_{p=1}^3 \sum_{l \in \mathcal{I}^{kj}_p} \alpha^k_l(u^k_l - \hat{\xi}^j_l) + \beta^k_l(\hat{\xi}^j_l - \ell^k_l) \leq v_j - c^0 \cdot x^k, \quad j \in [N], k \in [K],
$$

$$
\|C^\top x^k - \alpha^k + \beta^k\|_\infty \leq \lambda, \quad p \in [3], k \in [K]
$$
Plugging these reformulated constraints back into Problem (EC.51), we obtain the desired reformulation of the epigraph constraints. □