A data-driven approach to 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

We present a data-driven approach for solving multi-stage stochastic linear optimization problems in which uncertainty is correlated across stages. The proposed approach chooses decision rules which perform best when averaging over sample paths of a stochastic process; however, to avoid overfitting, we allow an adversary to slightly perturb each sample path. We show that this robust approach converges to the underlying stochastic problem as more data is obtained, even when the uncertainty is arbitrarily correlated across stages. Furthermore, we develop scalable approximation algorithms, which apply to problems with both continuous and integer decisions, by leveraging techniques from robust optimization. In computational experiments on stochastic inventory management problems, the proposed methods are practically tractable and produce decisions with near-optimal average performance. As a by-product of the aforementioned contributions, we also present new results and limitations of distributionally robust optimization with Wasserstein ambiguity sets.

Key words: Stochastic programming, Robust optimization, Sample-path approximations.

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. Linear optimization has, by all measures, succeeded as a framework for modeling and solving numerous real world problems. However, in many practical applications, the objective function and constraints are unknown at the time of decision making. To incorporate uncertainty into the linear optimization framework, Dantzig (1955) proposed partitioning the decision variables across multiple stages, which are made sequentially as more uncertain parameters are revealed. This formulation is known today as multi-stage linear optimization, which has become an integral modeling paradigm in many applications (e.g., supply chain management, energy planning, finance) and remains a focus of the stochastic and robust optimization communities (Birge and Louveaux 2011, Shapiro et al. 2009, Ben-Tal et al. 2009, Bertsimas et al. 2011).
In practice, decision makers increasingly have access to historical data which can provide valuable insight into future uncertainty. For example, consider a manufacturer which sells short lifecycle products. The manufacturer does not know a joint probability distribution of the demand over a new product’s lifecycle, but has access to historical demands over the lifecycle of similar products. Another example is energy planning, where operators must coordinate and commit to production levels throughout a day, the output of wind turbines is subject to uncertain weather conditions, and data on historical daily wind patterns is increasingly available. Other examples include portfolio management, where historical asset returns over time are available to investors, and transportation planning, where data comes in the form of historical ride usage of transit and ride sharing systems over the course of a day. Such historical data provides significant potential for operators to better understand how uncertainty unfolds through time, which can in turn be used for better planning.

When the underlying probability distribution is unknown, data-driven approaches to multi-stage stochastic linear optimization traditionally follow a two-step procedure. The historical data is first fit to a parametric model (e.g., an autoregressive moving average process), and decisions are then obtained by solving a multi-stage stochastic linear optimization problem using the estimated distribution. The estimation step is often essential, as techniques for solving multi-stage stochastic linear optimization (e.g., scenario tree discretization) generally require knowledge of the correlation structure of uncertainty across time; see Shapiro et al. (2009, Section 5.8). A fundamental difficulty in this approach is choosing a parametric model which will accurately estimate the underlying correlation structure and lead to good decisions.

Nonparametric data-driven approaches to multi-stage stochastic linear optimization where uncertainty is correlated across time are surprisingly scarce. Pflug and Pichler (2016) propose a nonparametric estimate-then-optimize approach based on applying a kernel density estimator to the historical data, which enjoys asymptotic optimality guarantees under a variety of strong technical conditions. Hanasusanto and Kuhn (2013) present another nonparametric approach wherein the conditional distributions in stochastic dynamic programming are estimated using kernel regression. Krokhmal and Uryasev (2007) discuss nonparametric path-grouping heuristics for constructing scenario trees from historical data. In the case of multi-stage stochastic linear optimization, to the best of our knowledge, there are no previous nonparametric data-driven approaches which are asymptotically optimal in the presence of time-dependent correlations. Importantly, in the absence of additional assumptions on the estimated distribution or on the problem setting, multi-stage stochastic linear optimization problems are well-known to be computationally demanding, suffering from the curse of dimensionality (Shapiro and Nemirovski 2005).
This paper aims to bring us closer to solving real world multi-stage problems directly from data. Our main contribution is a new data-driven approach for solving multi-stage stochastic linear optimization problems based on a simple robustification of the data. In particular, the proposed robust approach enjoys nonparametric guarantees on its average performance which, to the best of our knowledge, are the first of their kind for multi-stage stochastic linear optimization in data-driven settings. Moreover, leveraging algorithmic techniques from robust optimization, we show that the proposed approach is amenable to practically tractable approximations, even in problems with integer decisions, which can produce near-optimal solutions to multi-stage stochastic inventory problems with up to ten stages. As real-life applications typically start with historical data, the proposed methodology is readily applicable to settings regularly found in supply chain management, energy planning, and finance.

1.1. Related Literature

Our data-driven approach for multi-stage stochastic linear optimization draws on advances from robust optimization. We first discuss related results from the robust and distributionally robust optimization literature, followed by our key contributions.

Robust optimization has emerged as a prominent alternative to stochastic programming as a tractable framework for multi-stage decision making. 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 a probability distribution. Multi-stage robust linear optimization problems (also known as adjustable or adaptive optimization) are typically approximated by restricting the functional dependence of decisions on the uncertainty revealed up to that point; examples include linear decision rules (Ben-Tal et al. 2004), nonlinear functions via lifting (Chen and Zhang 2009), and finite adaptability (Bertsimas and Caramanis 2010, Postek and Hertog 2016, Bertsimas and Dunning 2016). For a comprehensive overview of decision rule approximation schemes, we refer the reader to Delage and Iancu (2015) and Georghiou et al. (2018). Several methods have also 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. 2018b). Despite the substantial progress in tractability and techniques for incorporating historical data, a central critique of robust optimization is that it does not aspire to find solutions which perform well on average.

Motivated by good average performance, the related model of distributionally robust optimization has recently received significant attention. First proposed by Scarf (1958), distributionally robust optimization models the uncertain parameters with a probability distribution, but the distribution is presumed to be 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, Wiesemann et al. 2014). Consequently, the extension of distributionally robust optimization to multi-stage decision making is an active area of research, including Bertsimas et al. (2018c) for multi-stage distributionally robust linear optimization with moment-based ambiguity sets. In addition, there has been a proliferation of data-driven constructions of ambiguity sets which emit probabilistic performance guarantees, including those based on the $p$-Wasserstein distance for $p \in [1, \infty)$ (Esfahani and Kuhn 2018, Fournier and Guillin 2015), phi-divergences (Ben-Tal et al. 2013, Bayraksan and Love 2015, Van Parys et al. 2017), and statistical hypothesis tests (Bertsimas et al. 2018a). Many of these data-driven approaches have since been applied to the particular case of two-stage distributionally robust linear optimization, including Jiang and Guan (2018) for phi-divergence and Hanasusanto and Kuhn (2018) for $p$-Wasserstein ambiguity sets when $p \in [1, \infty)$. To the best of our knowledge, no previous work has demonstrated whether such distributionally robust approaches, if extended to solve multi-stage stochastic linear optimization (with more than two stages) directly from data, retain their practical tractability and nonparametric performance guarantees.

1.2. Contributions

In this paper, we propose a new data-driven approach, based on robust optimization, for solving multi-stage stochastic linear optimization problems with unknown distributions. The approach provides nonparametric asymptotic optimality guarantees, providing assurance of its near-optimal approximation of the underlying stochastic problem in the presence of big data. At the same time, the approach remains computationally tractable and scalable by leveraging advances from robust optimization. The approach thus offers organizations across domains a general-purpose and practical tool for leveraging their historical data to find near-optimal operational decisions in dynamic environments. In greater detail, our key contributions are the following:

- We propose a new data-driven approach for solving multi-stage stochastic linear optimization problems with unknown distributions. The input is historical data in the form of sample paths of an underlying stochastic process. From this data, we optimize for the decision rules (i.e., policies which specify what decision to make in each stage as a function of the information revealed to that point) which perform best when averaging over the historical sample paths. However, to avoid overfitting, we add adversarial noise to each sample path. In contrast to robust optimization, the approach averages worst-case costs from multiple uncertainty sets, and thus aims to find decision rules with good average performance.
• We establish nonparametric convergence guarantees which show that the proposed data-driven approach provides a near-optimal approximation of multi-stage stochastic linear optimization in big data settings. Specifically, we prove under mild assumptions that the optimal cost and feasible decision rules for the proposed data-driven approach converge almost surely to those of the underlying multi-stage stochastic linear optimization problem as the number of sample paths tends to infinity. Importantly, these convergence guarantees hold for joint probability distributions with any correlation structure across the stages which, to the best of our knowledge, are first-of-kind results for multi-stage stochastic linear optimization in data-driven settings. The guarantees also do not require any restrictions on the space of decision rules and hold when decisions or random variables can be unbounded and mixed-integer.

• We show that the proposed approach is amenable to tractable approximations from robust optimization. Specifically, we extend linear decision rules and finite adaptability to obtain high quality solutions to the proposed data-driven framework in practical computation times. Through a new duality argument, we show that these approximation schemes can be reformulated as linear optimization problem which scales efficiently in the number of sample paths. The finite adaptability approach readily extends to problems with integer decision variables, and a practitioner can trade off the tightness of the approximation with an increase in computational cost.

• Under a particular construction of the uncertainty sets, we show that the proposed data-driven approach for multi-stage stochastic linear optimization can be interpreted as distributionally robust optimization using the $\infty$-Wasserstein ambiguity set. However, few previous convergence guarantees were known for this ambiguity set, even in single-stage settings (see discussion below). Thus, as a byproduct of the above convergence results, we show that distributionally robust optimization with the $\infty$-Wasserstein ambiguity set is asymptotically optimal under the same probabilistic assumptions as Esfahani and Kuhn (2018). In contrast, we show that possible alternatives based on other Wasserstein ambiguity sets do not generally exhibit asymptotic optimality when applied to multi-stage stochastic linear optimization.

• Across several examples from stochastic inventory management, we demonstrate that the proposed data-driven methodology finds near-optimal solutions in practical computation times, significantly outperforming previous approaches. Specifically, we first consider a multi-stage stochastic inventory problem with autoregressive demands. Second, we consider a multi-stage mixed-integer stochastic lot-sizing problem for short lifecycle products. In both cases, the demands are correlated across time, and our only information comes from historical data. Using
the proposed data-driven approaches, we obtain decision rules with near-optimal average out-of-sample performance in minutes for problems with up to ten stages and hundreds of sample paths.

To the best of our knowledge, our proposed data-driven approach of averaging over multiple uncertainty sets is novel in application to multi-stage stochastic linear optimization. Xu et al. (2012) considered averaging over multiple uncertainty sets to establish convergence guarantees for predictive machine learning methods, drawing connections with distributionally robust optimization and kernel density estimation. Their convergence results require that the objective function is continuous, the underlying distribution is continuous, and there are no constraints on the support. Absent strong assumptions on the problem setting and on the space of decision rules (which in general can be discontinuous), these properties do not hold in multi-stage problems. Erdoğan and Iyengar (2006) provide feasibility guarantees on robust constraints over unions of uncertainty sets with the goal of approximating ambiguous chance constraints using the Prohorov metric. Their probabilistic guarantees require that the constraint functions have a finite VC-dimension (Erdoğan and Iyengar 2006, Theorem 5), an assumption which does not hold in general for two- or multi-stage problems (Erdoğan and Iyengar 2007).

Our convergence guarantees are also novel in the context of Wasserstein-based distributionally robust optimization. Under a particular construction of the uncertainty sets, we show that the proposed data-driven approach to multi-stage stochastic linear optimization can be interpreted as distributionally robust optimization using the $\infty$-Wasserstein ambiguity set. However, the convergence results in our paper do not make use of this interpretation, as there are surprisingly few previous convergence results for this ambiguity set, even in single-stage settings. Indeed, when an underlying distribution is unbounded, the $\infty$-Wasserstein distance between an empirical distribution and true distribution is always infinite (Givens and Shortt 1984) and thus does not converge to zero as more data is obtained. Therefore, it is not possible to develop measure concentration guarantees for the $\infty$-Wasserstein distance (akin to those of Fournier and Guillin (2015)) which hold in general for light-tailed but unbounded probability distributions. Consequently, the proof techniques used by Esfahani and Kuhn (2018, Theorem 3.6) to establish convergence guarantees for the 1-Wasserstein ambiguity set do not appear to extend to the $\infty$-Wasserstein ambiguity set. In this paper, we establish general convergence guarantees for the proposed data-driven approach by developing connections with nonparametric support estimation (Devroye and Wise 1980) and new bounds for distributionally robust optimization with the 1-Wasserstein ambiguity set. Consequently, the results in this paper actually imply general convergence guarantees for distributionally
robust optimization with the $\infty$-Wasserstein ambiguity set under mild probabilistic assumptions for the first time.

Our paper is organized as follows. Section 2 introduces multi-stage stochastic linear optimization a the data-driven setting. Section 3 presents the new data-driven framework. Section 4 states the main convergence guarantees. Section 5 presents approximation algorithms for the proposed data-driven framework based on linear decision rules and finite adaptability. Section 6 discusses the relationships between the proposed data-driven framework and distributionally robust optimization. Section 7 demonstrates the performance and tractability of the proposed algorithm in computational experiments. Section 8 offers concluding thoughts. All technical proofs are relegated to the attached appendices.

**Notation.** We denote the real numbers by $\mathbb{R}$, the nonnegative real numbers by $\mathbb{R}_+$, the integers by $\mathbb{Z}$. Lowercase and uppercase bold letters refer to vectors and matrices (e.g., $\mathbf{v} \in \mathbb{R}^d$ and $\mathbf{M} \in \mathbb{Z}^{d \times k}$). We assume throughout that $\| \cdot \|$ refers to an $\ell_p$-norm in $\mathbb{R}^d$ (such as $\| \mathbf{v} \|_1 = \sum_{i=1}^d |v_i|$ or $\| \mathbf{v} \|_\infty = \max_{i \in [d]} |v_i|$). We let $\emptyset$ denote the empty set, $\text{int}(P)$ be the interior of a set $P \subseteq \mathbb{R}^d$, and $[K]$ be shorthand for the set of consecutive integers $\{1, \ldots, K\}$. Throughout the paper, we let $\xi := (\xi_1, \ldots, \xi_T) \in \mathbb{R}^d$ denote a stochastic process with a joint probability distribution $\mathbb{P}$ and assume that $\hat{\xi}_1, \ldots, \hat{\xi}_N$ are independent and identically distributed (i.i.d.) samples from that distribution. Let $\mathbb{P}^N := \mathbb{P} \times \cdots \times \mathbb{P}$ denote the $N$-fold probability distribution over the historical data, and let $S \subseteq \Xi$ denote the support of the true distribution; that is, the smallest closed set where $\mathbb{P}(\xi \in S) = 1$. The extended real numbers are defined as $\bar{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$, and we adopt the convention that $\infty - \infty = \infty$. Furthermore, the expectation of an extended-valued function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ applied to the stochastic process is denoted by $\mathbb{E}[f(\xi)] = \mathbb{E}_{\mathbb{P}}[f(\xi)] = \mathbb{E}_{\mathbb{P}}[\max\{f(\xi), 0\}] - \mathbb{E}_{\mathbb{P}}[\max\{-f(\xi), 0\}]$. Therefore, it is always the case that either $\mathbb{E}[f(\xi)] = \infty$, $\mathbb{E}[f(\xi)] = -\infty$, or $\mathbb{E}[|f(\xi)|] < \infty$, and the strong law of large numbers holds in each of those cases (Williams 1991, Theorem 12.10). Finally, for any set $\mathcal{Z} \subseteq \mathbb{R}^d$, we let $\mathcal{P}(\mathcal{Z})$ denote the set of all probability distributions on $\mathbb{R}^d$ which satisfy $\mathbb{Q}(\xi \in \mathcal{Z}) \equiv \mathbb{E}_{\mathbb{Q}}[I\{\xi \in \mathcal{Z}\}] = 1$.

2. Problem Setting

We consider multi-stage stochastic linear optimization problems with $T \geq 1$ stages. The uncertain parameters observed over the time horizon are represented by a stochastic process $\xi := (\xi_1, \ldots, \xi_T) \in \mathbb{R}^d$, where $\xi_t \in \mathbb{R}^{d_t}$ is a random variable that is observed immediately after the decision in stage $t$ is selected. A decision rule $\mathbf{x} := (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ is a collection of policies which specify what decision to make in each stage based on the information observed up to that point. More precisely, the policy in each stage is a function of the form $\mathbf{x}_t : \mathbb{R}^{d_1} \times \cdots \times \mathbb{R}^{d_{t-1}} \to \mathcal{X}_t$ where the set $\mathcal{X}_t$ enforces which of
the decision components are continuous or integer. We use the shorthand notation \( x \in \mathcal{X} \) to denote such decision rules.

In multi-stage stochastic linear optimization, our goal is to find a decision rule which minimizes a linear cost function in expectation while satisfying a system of linear inequalities almost surely. These problems are represented by

\[
\begin{align*}
\text{minimize} \quad & \mathbb{E} \left[ \sum_{t=1}^{T} c_t(\xi) \cdot x_t(\xi_1, \ldots, \xi_{t-1}) \right] \\
\text{subject to} \quad & \sum_{t=1}^{T} A_t(\xi) x_t(\xi_1, \ldots, \xi_{t-1}) \leq b(\xi) \quad \text{a.s.}
\end{align*}
\]

(1)

Following standard convention, we assume that the problem parameters \( c_1(\xi) \in \mathbb{R}^{n_1}, \ldots, c_T(\xi) \in \mathbb{R}^{n_T}, A_1(\xi) \in \mathbb{R}^{m \times n_1}, \ldots, A_T(\xi) \in \mathbb{R}^{m \times n_T}, \) and \( b(\xi) \in \mathbb{R}^{m} \) are linear functions of the stochastic process. Numerous applications across various disciplines are modeled in this framework, and examples are found in Section 7 and Birge and Louveaux (2011).

In this paper, we assume that the underlying probability distribution of the stochastic process is unknown. Instead, our information comes from historical data of the form

\[ \hat{\xi}_j \equiv (\hat{\xi}_j^1, \ldots, \hat{\xi}_j^T), \quad j = 1, \ldots, N, \]

and we refer to each of these trajectories as a sample path of the stochastic process. This setting corresponds to many real-life applications. For example, consider managing the inventory of a new short lifecycle product, in which production decisions must be made over the product’s lifecycle. In this case, each sample path represents the historical sales data observed over the lifecycle of a comparable product. Further examples are readily found in energy planning and finance, among many others. Our goal in this paper is a general-purpose, data-driven approach which produces near-optimal decision rules to Problem (1) in practical computation times.

We will also assume that the support of the probability distribution is unknown. For example, in inventory management, an upper bound on the demand, if one exists, is generally unknown. On the other hand, we often have partial knowledge on the underlying support. For example, when the stochastic process captures the demand for a new product or the energy produced by a wind turbine, it is often the case that the uncertainty will be nonnegative. To allow any partial knowledge on the support to be incorporated, we assume knowledge of a convex superset \( \Xi \subseteq \mathbb{R}^d \) of the support of the underlying joint distribution, that is, \( \mathbb{P}(\xi \in \Xi) = 1 \).
3. A Robust Approach to Multi-Stage Stochastic Linear Optimization

In this section, we present the proposed data-driven approach, based on robust optimization, for solving multi-stage stochastic linear optimization. First, we construct an uncertainty set $U_N^j \subseteq \Xi$ around each sample path, consisting of realizations $\zeta \equiv (\zeta_1, \ldots, \zeta_T)$ which are slight perturbations of $\hat{\xi}^j \equiv (\hat{\xi}_1^j, \ldots, \hat{\xi}_T^j)$. Then, we optimize for 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:

$$\minimize_{x \in X} \frac{1}{N} \sum_{j=1}^{N} \sup_{\zeta \in U_N^j} \sum_{t=1}^{T} c_t(\zeta) \cdot x_t(\zeta_1, \ldots, \zeta_{t-1})$$

subject to $\sum_{t=1}^{T} A_t(\zeta) x_t(\zeta_1, \ldots, \zeta_{t-1}) \leq b(\zeta) \quad \forall \zeta \in \bigcup_{j=1}^{N} U_N^j.$

(2)

In contrast to traditional robust optimization, Problem (2) involves averaging 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. We note that Problem (2) only requires that the decision rules are feasible for the realizations in the uncertainty sets. These feasibility requirements are justified when the overlapping uncertainty sets encompasses the variability of future realizations of the uncertainty; see Section 4.

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

$$U_N^j := \{ \xi = (\xi_1, \ldots, \xi_T) \in \Xi : \| \xi - \hat{\xi}^j \| \leq \epsilon_N \},$$

where $\epsilon_N \geq 0$ is a parameter which controls the size of the uncertainty sets. The rationale for this particular uncertainty set is three-fold. First, it is conceptually simple, requiring only a single parameter to both estimate the expectation in the objective and the support of the distribution in the constraints. Second, under appropriate choice of the robustness parameter, we will show that Problem (2) with these uncertainty sets has desirable nonparametric out-of-sample performance guarantees for Problem (1) (see Section 4). Finally, the uncertainty sets are of similar structure, which can be exploited to obtain tractable reformulations (see Section 5).

Our approach, in a nutshell, uses robust optimization as a tool for solving multi-stage stochastic linear optimization directly from data. More specifically, we obtain decision rules and estimate the optimal cost of Problem (1) by solving Problem (2). We refer the proposed data-driven approach for solving multi-stage stochastic linear optimization problems as sample or sample-path robust optimization. As mentioned previously, the purpose of robustness is to ensure that resulting decision
rules do not overfit the historical sample paths. To illustrate this role performed by robustness, we consider the following example.

**Example 1.** Consider a supplier which aims to satisfy uncertain demand over two phases at minimal cost. The supplier selects an initial production quantity at $1 per unit after observing preorders, and produces additional units at $2 per unit after the regular orders are received. To determine the optimal production levels, we wish to solve

\[
\begin{align*}
\text{minimize} & \quad \mathbb{E} [x_2(\xi_1) + 2x_3(\xi_1, \xi_2)] \\
\text{subject to} & \quad x_2(\xi_1) + x_3(\xi_1, \xi_2) \geq \xi_1 + \xi_2 \quad \text{a.s.} \\
& \quad x_2(\xi_1), x_3(\xi_1, \xi_2) \geq 0 \quad \text{a.s.}
\end{align*}
\]

(3)

The output of the optimization problem are policies, \(x_2 : \mathbb{R} \to \mathbb{R}\) and \(x_3 : \mathbb{R}^2 \to \mathbb{R}\), which specify what production levels to choose as a function of the demands observed up to that point. The joint probability distribution of the demand process \((\xi_1, \xi_2) \in \mathbb{R}^2\) is unknown, and the supplier’s knowledge comes from historical demand realizations of past products, denoted by \((\hat{\xi}_1, \hat{\xi}_2), \ldots, (\hat{\xi}_N^1, \hat{\xi}_N^2)\).

For the sake of illustration, suppose we attempted to approximate Problem (3) by choosing the decision rules which perform best when averaging over the historical data without any robustness. Such a sample average approach amounts to solving

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{N} \sum_{j=1}^{N} \left( x_2(\hat{\xi}_j^1) + 2x_3(\hat{\xi}_j^1, \hat{\xi}_j^2) \right) \\
\text{subject to} & \quad x_2(\hat{\xi}_j^1) + x_3(\hat{\xi}_j^1, \hat{\xi}_j^2) \geq \hat{\xi}_j^1 + \hat{\xi}_j^2 \quad \forall j \in [N] \\
& \quad x_2(\hat{\xi}_j^1), x_3(\hat{\xi}_j^1, \hat{\xi}_j^2) \geq 0 \quad \forall j \in [N].
\end{align*}
\]

Suppose that the random variable \(\xi_1\) for preorders has a continuous distribution. In that case, it immediately follows that \(\hat{\xi}_1^1 \neq \cdots \neq \hat{\xi}_N^1\) almost surely, and thus an optimal decision rule for the above optimization problem is

\[
x_2(\xi_1) = \begin{cases} 
\hat{\xi}_j^1 + \hat{\xi}_j^2, & \text{if } \xi_1 = \hat{\xi}_j^1 \text{ for } j \in [N], \\
0, & \text{otherwise;}
\end{cases}
\]

\[
x_3(\xi_1, \xi_2) = 0.
\]

Unfortunately, these decision rules are *nonsensical* with respect to Problem (3). Indeed, the decision rules will not result in feasible decisions for the true stochastic problem with probability one. Moreover, the optimal cost of the above optimization problem will converge almost surely to \(\mathbb{E} [\xi_1 + \xi_2]\) as the number of sample paths \(N\) tends to infinity, which can in general be far from that of the stochastic problem. Clearly, such a sample average approach results in overfitting, even in big data settings, and thus provides an unsuitable approximation of Problem (3). \(\square\)
A key takeaway from this paper is that this overfitting phenomenon in the above example is eliminated by adding robustness to the historical data. In particular, we will show in the following section that when the robustness parameter $\epsilon_N$ is chosen correctly, Problem (2) converges to a near-optimal approximation of Problem (1) as more data is obtained, without requiring any parametric assumptions on the stochastic process or restrictions on the space of decision rules.

4. Convergence Guarantees

In this section, we present theoretical guarantees showing that Problem (2) provides a near-optimal approximation of multi-stage stochastic linear optimization in big data settings. More specifically, under an appropriate construction of the uncertainty sets, we show that the feasible decision rules and optimal cost of Problem (2) converge almost surely to those of Problem (1) as more data is obtained. In Section 4.1, we state and explain our assumptions used in the subsequent convergence results. In Section 4.2, we present our feasibility guarantees. In Section 4.3, we establish asymptotic optimality for the data-driven approach. The practical value of these convergence guarantees are illustrated via computational experiments in Section 7.

4.1. Assumptions

We begin by introducing our assumptions which will be used for establishing convergence guarantees. First, we will assume that the joint probability distribution of the stochastic process satisfies the following light-tail assumption:

\textbf{Assumption 1.} There exists a constant $a > 1$ such that $b := \mathbb{E} \left[ \exp \left( \| \xi \|^a \right) \right] < \infty$.

For example, Assumption 1 is satisfied when the stochastic process has a multivariate Gaussian distribution. Importantly, this assumption does not require any parametric assumptions on the correlation structure of the random variables across stages, and we do not assume that the coefficient $a > 1$ is known.

Second, we will assume that the robustness parameter $\epsilon_N$ is chosen to be strictly positive and decreases to zero as more data is obtained at the following rate:

\textbf{Assumption 2.} There exists a constant $\kappa > 0$ such that $\epsilon_N := \kappa N^{-\max\{3, d+1\}}$.

In a nutshell, Assumption 2 provides a theoretical requirement on the robustness parameter to ensure that Problem (2) will not overfit the historical data (see Example 1 from Section 3) and provides practical guidance on how the robustness parameter can be updated as more data is obtained. We note that, for many of the following results, the robustness parameter can decrease to zero at a faster rate; nonetheless, we shall impose Assumption 2 for all our results for simplicity.
Finally, our convergence guarantees for Problem (2) do not require any restrictions on the space of decision rules. Our analysis will only require the following mild assumption on the problem structure.

**Assumption 3.** There exists a $L \geq 0$ such that, for all $N \in \mathbb{N}$, the optimal cost of Problem (2) would not change if we added the following constraints:

$$
\sup_{\zeta \in \bigcup_{j=1}^{N} U_j^N} \| x_t(\zeta_1, \ldots, \zeta_{t-1}) \| \leq \sup_{\zeta \in \bigcup_{j=1}^{N} U_j^N} L (1 + \| \zeta \| ) \quad \forall t \in [T].
$$

In other words, this assumption says that there is always a near-optimal decision rule to Problem (2) where the decisions which result from realizations in uncertainty sets are bounded by the largest realization in the uncertainty sets. Moreover, this is a mild assumption that is easily verifiable in many practical examples. Indeed, in Appendix A, we show that every example presented in this paper satisfies this assumption.

### 4.2. Feasibility guarantees

We begin by discussing the out-of-sample feasibility of decision rules for Problem (2). Recall that Problem (2) finds decision rules which are feasible for each realization in the uncertainty sets. However, one cannot guarantee that these decision rules will be feasible for realizations outside of the uncertainty sets. Thus, a pertinent question is whether a decision rule obtained from approximately solving Problem (2) is feasible with high probability. To address the question of feasibility, we propose leveraging classic results from detection theory.

Let $S_N := \bigcup_{j=1}^{N} U_j^N$ be shorthand for the union of the uncertainty sets. We say that a decision rule is $S_N$-feasible if

$$
\sum_{i=1}^{T} A_i(\zeta) x_i(\zeta_1, \ldots, \zeta_{t-1}) \leq b(\zeta) \quad \forall \zeta \in S_N.
$$

In other words, the set of feasible decision rules to Problem (2) are exactly those which are $S_N$-feasible. Our subsequent analysis utilizes the following (seemingly tautological) observation: for any decision rule that is $S_N$-feasible, we have that

$$
\mathbb{P} \left( \sum_{i=1}^{T} A_i(\zeta) x_i(\zeta_1, \ldots, \zeta_{t-1}) \leq b(\zeta) \right) \geq \mathbb{P}(\zeta \in S_N),
$$

where $\mathbb{P}(\zeta \in S_N)$ is shorthand for $\mathbb{P}(\zeta \in S_N \mid \hat{\zeta}_1, \ldots, \hat{\zeta}_N)$. 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 $\zeta \in S_N$. 
We have thus transformed the analysis of feasible decision rules for Problem (2) to the problem of analyzing the performance of $S_N$ as an estimate of the support $S$ of the stochastic process. Interestingly, this nonparametric estimator for the support of a joint probability 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 nonparametric estimate of $S$ has been studied with applications in cluster analysis and image recognition (Korostelev and Tsybakov 1993, Schölkopf et al. 2001). Leveraging this new connection between stochastic optimization and support estimation, we obtain the following guarantee on feasibility.

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

$$
\lim_{N \to \infty} \left( \frac{N^{1/d}}{(\log N)^{d+1}} \right) \mathbb{P}(\xi \notin S_N) = 0.
$$

**Proof.** See Appendix B. □

Intuitively speaking, Theorem 1 provides a guarantee that any feasible decision rule to Problem (2) will be feasible with high probability on future data when the number of sample paths is large. To illustrate why robustness is indeed necessary to achieve such feasibility guarantees, we recall from Example 1 that decision rules may prohibitively overfit the data and be infeasible with probability one if the parameter $\epsilon_N$ is set to zero.

### 4.3. Asymptotic optimality

We now shift focus to analyzing the optimal cost of Problem (2). Specifically, we shall show that the optimal cost of Problem (2) nearly converges almost surely to the optimal cost of the true multi-stage stochastic linear optimization problem, providing assurance that Problem (2) will result in near-optimal average performance in big data settings. For notational convenience, let $J^*$ be the optimal cost of Problem (1), $\tilde{J}_N$ be the optimal cost of Problem (2), and $S \subseteq \Xi$ be the support of the underlying joint probability distribution of the stochastic process.

Our main result presents tight asymptotic lower and upper bounds on the optimal cost $\tilde{J}_N$ of Problem (2). First, let $J$ be defined as the maximal optimal cost of any chance-constrained variant of the multi-stage stochastic linear optimization problem:

$$
J := \lim_{\rho \downarrow 0} \min_{x \in X, S \subseteq \Xi} \mathbb{E} \left[ \sum_{t=1}^{T} c_t(\xi) \cdot x_t(\xi_1, \ldots, \xi_{t-1}) I\{\xi \in \tilde{S}\} \right]
$$

subject to

$$
\sum_{t=1}^{T} A_t(\zeta)x_t(\zeta_1, \ldots, \zeta_{t-1}) \leq b(\zeta) \quad \forall \zeta \in \tilde{S}
$$

$$
\mathbb{P}(\xi \in \tilde{S}) \geq 1 - \rho.
$$
Note that the above limit must exist, as the optimal cost of the above chance constrained optimization problem is monotone in \( \rho \). We observe that \( J \) is lower bound on \( J^* \), since there is a feasible solution can always be constructed where \( \tilde{S} \) is any set which contains the stochastic process almost surely. Nonetheless, \( J \) and \( J^* \) can often be equal, as is the case for the stochastic inventory management problem in Example 1 from Section 3 (see Appendix C).

Second, let \( \bar{J} \) be the optimal cost of the multi-stage stochastic linear optimization problem with an additional restriction that the decision rules are feasible on an expanded support:

\[
\bar{J} := \lim_{\rho \downarrow 0} \min_{x \in X} \mathbb{E} \left[ \sum_{t=1}^{T} c_t(\xi) \cdot x_t(\xi_1, \ldots, \xi_{t-1}) \right]
\]

subject to \( \sum_{t=1}^{T} A_t(\zeta)x_t(\zeta_1, \ldots, \zeta_{t-1}) \leq b(\zeta) \quad \forall \zeta \in \mathcal{F} : \text{dist}(\zeta, S) \leq \rho. \)

We remark that the limit as \( \rho \) tends down to zero must exist as well, since the optimal cost 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 \( \mathbb{E}[\cdot] \), which we define here as the local upper semicontinuous envelope of an expectation, \( \mathbb{E}[f(\xi)] := \lim_{\epsilon \to 0} \mathbb{E}[\sup_{\zeta \in \mathcal{F} : \|\zeta - \xi\| \leq \epsilon} f(\xi)]. \) We similarly observe that \( \bar{J} \) is an upper bound on \( J^* \), since the above optimization problem involves additional constraints and an upper envelope of the objective function.

Our main result is the following:

**THEOREM 2.** Suppose Assumptions 1, 2, and 3 hold. Then, \( \mathbb{P}^\infty \)-almost surely we have

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

**Proof.** See Appendix D. \( \square \)

Theorem 2 provides assurance that the proposed data-driven framework becomes a near-optimal approximation of multi-stage stochastic linear optimization. Note that Theorem 2 holds in very general cases; for example, it does not require boundedness on the decisions or random variables, requires no parametric assumptions on the correlations across stages, and holds when the decisions contain both continuous and integer components. Moreover, the convergence of Problem (2) does not necessitate imposing any restrictions on the space of decision rules. To the best of our knowledge, such nonparametric asymptotic optimality guarantees are first of their kind for data-driven multi-stage stochastic linear optimization when uncertainty is correlated across stages.

In many practical applications, we do not expect the gap between the lower and upper bounds to be large. Nonetheless, in the absence of additional assumptions, the bounds offered in Theorem 2 appear to be tight, as illustrated by the following examples:
Example 2. Consider the following stochastic optimization problem:

$$\begin{align*}
\text{minimize} & \quad x_1 \\
\text{subject to} & \quad x_1 \geq \xi_1 \quad \text{a.s.}
\end{align*}$$

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 the sample robust optimization approach to the above optimization problem with a robustness parameter $\epsilon_N = N^{-\frac{1}{3}}$. Suppose the true probability distribution is defined by $\mathbb{P}(\xi_1 > \alpha) = (1 - \alpha)^k$ for some fixed $k > 0$ and every $\alpha \in [0, 1]$. In this case, the true support of the underlying distribution is $[0, 1]$ for any $k > 0$. Then, employing the Hewitt-Savage zero-one law, we prove in Appendix E that the bounds in Theorem 2 are tight under different choices of $k$:

<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>

More generally, this example illustrates that the bounds from Theorem 2 can hold at equality or strict inequality when the feasibility of decisions depends on future uncertainty. □

5. Tractable Approximations

In the previous section, we showed that Problem (2) provides a near-optimal approximation of multi-stage stochastic linear optimization in big data settings. In this section, we demonstrate that Problem (2) can be tractably solved using approximation schemes from robust optimization. Specifically, we show that the techniques of linear decision rules and finite adaptability from robust optimization can be extended to obtain high-quality decision rules to Problem (2) in practical computation times, including for problems with continuous and integer decisions. In particular, we present a simple duality technique (Theorem 3) which allows the computational cost of these algorithms to scale efficiently in the number of sample paths, thus remaining practical in big data settings. The computational tractability and solution quality of these approximation schemes on problems with hundreds of sample paths is demonstrated via empirical experiments in Section 7.
5.1. Linear decision rules

Generally speaking, multi-stage optimization problems are computationally demanding due to optimizing over an unrestricted space of decision rules. To overcome this challenge, a common approximation technique in robust optimization is to restrict the space of decision rules to a space which can more easily be optimized. As described in the introduction, the success of robust optimization as a modeling framework for addressing real world multi-stage problems is often attributed the computational tractability of such decision rule approximations. This section extends one such decision rule scheme, known as linear decision rules, for approximately solving Problem (2) and illustrates its computational tractability in big data settings.

Specifically, we consider approximating Problem (2) by restricting its decision rules to those of the form

\[ x_t(\zeta_1, \ldots, \zeta_{t-1}) = x_{t,0} + \sum_{s=1}^{t-1} X_{t,s} \zeta_s. \]

Thus, rather than optimizing over a space of decision rules (functions), we instead optimize over a finite collection of decision variables which define the linear decision rule. For the setting where \( c_t(\xi) \) and \( A_t(\xi) \) do not depend on the uncertain parameters and all decision variables are continuous, the resulting linear decision rule approximation of Problem (2) is given by

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{N} \sum_{j=1}^{N} \sup_{\zeta \in U_j} \sum_{t=1}^{T} c_t \cdot \left( x_{t,0} + \sum_{s=1}^{t-1} X_{t,s} \zeta_s \right) \\
\text{subject to} & \quad \sum_{t=1}^{T} A_t \left( x_{t,0} + \sum_{s=1}^{t-1} X_{t,s} \zeta_s \right) \leq b(\zeta) \quad \forall \zeta \in \bigcup_{j=1}^{N} U_j,
\end{align*}
\]

(4)

where the decision variables are \( x_{t,0} \in \mathbb{R}^{n_t} \) and \( X_{t,s} \in \mathbb{R}^{n_t \times d_s} \) for all \( 1 \leq s < t \leq T \) and the linear function \( b(\zeta) \in \mathbb{R}^m \) is shorthand for \( b^0 + \sum_{t=1}^{T} B_t \zeta_t \).

Much like linear decision rules in robust optimization, we observe that Problem (4), when feasible, always produces a feasible solution for Problem (2) and an upper bound on its optimal cost. Nonetheless, Problem (4) has semi-infinite constraints, which must be eliminated in order for the optimization problem to be solvable by off-the-shelf solvers. A standard technique from robust optimization for eliminating semi-infinite constraints is to introduce (dual) auxiliary decision variables and constraints for each uncertainty set. Importantly, for Problem (4) be practically tractable in the presence of big data, the size of an equivalent finite-dimensional optimization problem must scale efficiently in the number of sample paths.

We now show that Problem (4) can be reformulated as a tractable linear optimization with size that scales lightly in the number of sample paths (Theorem 3). The central idea enabling the
following reformulation is that the worst-case realizations over the various uncertainty sets are found by optimizing over identical linear functions. Thus, when constructing the robust counterparts for each uncertainty set, we can combine the dual auxiliary decision variables from different uncertainty sets, resulting in a reformulation where the number of auxiliary decision variables is independent of the number of sample paths. To illustrate this reformulation technique, we focus on uncertainty sets which satisfy the following construction:

**Assumption 4.** The uncertainty sets have the form

\[ U^j_N := \{ \zeta \in \mathbb{R}^d : \ell^j \leq \zeta \leq u^j \} \]

For example, Assumption 4 holds if we choose the set \( \Xi \) to be \( \mathbb{R}^d_+ \) and use the \( \| \cdot \|_\infty \) norm in the uncertainty sets from Section 3. Our main result, which illustrates the novel duality technique described above, is the following:

**Theorem 3.** If Assumption 4 holds, then Problem (4) can be reformulated as a linear optimization problem with \( O(md) \) auxiliary decision variables and \( O(md + mN) \) linear constraints.

**Proof.** By introducing epigraph variables \( v_1, \ldots, v_N \in \mathbb{R} \), the constraints in Problem (4) can be rewritten as

\[
\begin{align*}
\sum_{t=1}^{T} \left( \sum_{s=t+1}^{T} X^t_s c_s \right) \cdot \zeta_t & \leq v_j - \sum_{t=1}^{T} c_t \cdot x_{t,0} \quad \forall \zeta \in U^j_N, j \in \{1, \ldots, N\}, \\
\sum_{t=1}^{T} \left( -B_t + \sum_{s=t+1}^{T} A_s X^t_{s,s} \right) \cdot \zeta_t & \leq b^0 - \sum_{t=1}^{T} A_t x_{t,0} \quad \forall \zeta \in U^j_N, j \in \{1, \ldots, N\}.
\end{align*}
\]

We will now reformulate each of these semi-infinite constraints by introducing auxiliary variables. First, we observe that each of the above semi-infinite constraints can be rewritten as

\[
\max_{\zeta \in U^j_N} \sum_{t=1}^{T} d_t \cdot \zeta_t \leq \gamma
\]

for some vector \( d := (d_1, \ldots, d_T) \in \mathbb{R}^d \) and scalar \( \gamma \in \mathbb{R} \). Moreover, it follows from strong duality for linear optimization that

\[
\max_{\zeta \in U^j_N} \sum_{t=1}^{T} d_t \cdot \zeta_t = \begin{cases} 
\text{minimize} & \sum_{t=1}^{T} (u^j_t \cdot \mu_t - \ell^j_t \cdot \lambda_t) \\
\text{subject to} & \mu_t - \lambda_t = d_t \quad \forall t \in [T],
\end{cases}
\]

where \( u^j := (u^j_1, \ldots, u^j_T) \in \mathbb{R}^d \) and \( \ell^j := (\ell^j_1, \ldots, \ell^j_T) \in \mathbb{R}^d \) are the upper and lower bounds which define the uncertainty set. We readily observe that the solutions \( \mu_t = [d_t]_+ \) and \( \lambda_t = [-d_t]_+ \) are optimal for the above optimization problem. Importantly, these optimal solutions to the dual problem are independent of the index \( j \). Thus, the semi-infinite constraints in the epigraph formulation of
Problem (4) are satisfied if and only if there exists \( \alpha := (\alpha_1, \ldots, \alpha_T) \in \mathbb{R}^d_+ \) and \( \beta := (\beta_1, \ldots, \beta_T) \in \mathbb{R}^d_+ \) which satisfy
\[
\sum_{t=1}^{T} (\alpha_t \cdot u^t_j - \beta_t \cdot \ell^t_j + c_t \cdot x_{t,0}) \leq v_j \quad \forall j \in [N]
\]
\[
\alpha_t - \beta_t = \sum_{s=t+1}^{T} X_{s,t}^\top c_s \quad \forall t \in [T]
\]
and there exists \( M := (M_1, \ldots, M_T) \in \mathbb{R}^{m \times d}_+ \) and \( \Lambda := (\Lambda_1, \ldots, \Lambda_T) \in \mathbb{R}^{m \times d}_+ \) which satisfy
\[
\sum_{t=1}^{T} (M_t u^t_j - \Lambda_t \ell^t_j + A_t x_{t,0}) \leq b^0 \quad \forall j \in [N]
\]
\[
M_t - \Lambda_t = -B_t + \sum_{s=t+1}^{T} A_s X_{t,s} \quad \forall t \in [T]
\]
Removing the epigraph decision variables, the resulting reformulation of Problem (4) is
\[
\begin{align*}
\text{minimize} & \quad \frac{1}{N} \sum_{j=1}^{N} \sum_{t=1}^{T} (\alpha_t \cdot u^t_j - \beta_t \cdot \ell^t_j + c_t \cdot x_{t,0}) \\
\text{subject to} & \quad \alpha_t - \beta_t = \sum_{s=t+1}^{T} X_{s,t}^\top c_s \quad t \in [T] \\
& \quad \sum_{t=1}^{T} (M_t u^t_j - \Lambda_t \ell^t_j + A_t x_{t,0}) \leq b^0 \quad j \in [N] \\
& \quad M_t - \Lambda_t = -B_t + \sum_{s=t+1}^{T} A_s X_{t,s} \quad t \in [T]
\end{align*}
\]
where the auxiliary decision variables are \( \alpha \equiv (\alpha_1, \ldots, \alpha_T), \beta \equiv (\beta_1, \ldots, \beta_T) \in \mathbb{R}^d_+ \) and \( M \equiv (M_1, \ldots, M_T), \Lambda \equiv (\Lambda_1, \ldots, \Lambda_T) \in \mathbb{R}^{m \times d}_+ \). \( \Box \)

While linear decision rules can sometimes provide a near-optimal approximation of Problem (2) (see Section 7.1), we do not expect this to be the case in general. Indeed, we recall from Section 4 that Problem (2) becomes a near-optimal approximation of Problem (1) in the presence of big data, and it has been known from the early literature that linear decision rules generally provide a poor approximation for multi-stage stochastic linear optimization; see, e.g., Garstka and Wets (1974, Section 6). Furthermore, linear decision rules do not allow for adaptability for problems with integer decision variables. Nonetheless, we can obtain tighter approximations of Problem (2) by selecting a richer space of decision rules, an abundance of which can be found in the robust optimization literature. Moreover, the scalable reformulation techniques used in Theorem 3 can generally be extended to richer classes of decision rules, such as finite adaptability (see Section 5.2). Finally, Problem (2) is also amenable to new approximation schemes which exploit its particular structure;
we refer to our companion paper Bertsimas et al. (2019) for such an approximation algorithm for two-stage problems. In all cases, and as a result of the convergence guarantees from Section 4, Problem (2) offers an opportunity to extend algorithmic advances from robust optimization to obtain new approximations of multi-stage stochastic linear optimization.

5.2. Finite adaptability

In this section, we show how to extend the decision rule approximation scheme of finite adaptability from robust optimization (Bertsimas and Caramanis 2010) to obtain tighter approximations of Problem (2). Specifically, finite adaptability allows the set \( \Xi \) to be partitioned into smaller regions, and a separate static or linear decision rule is optimized in each region. The approach of finite adaptability 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 show that the duality techniques from the previous section (Theorem 3) readily extend to this richer class of decision rules, and the practical performance of this approach is demonstrated in Section 7.

We begin by describing the approximation scheme of finite adaptability from robust optimization. In finite adaptability, one partitions the uncertainty set into different regions, and optimizes a separate linear decision rule for each region. Let \( P^1, \ldots, P^K \subseteq \mathbb{R}^d \) be regions which form a partition of \( \Xi \subseteq \mathbb{R}^d \). For each stage \( t \), let \( P^k_t \subseteq \mathbb{R}^{d_1 + \cdots + d_t} \) be the projection of the region \( P^k \) onto the first \( t \) stages. Then, we consider approximating Problem (2) by restricting its decision rules to those of the form

\[
x_t(\zeta_1, \ldots, \zeta_{t-1}) = \begin{cases} 
  x^1_{t,0} + \sum_{s=1}^{t-1} X^1_{t,s} \zeta_s, & \text{if } (\zeta_1, \ldots, \zeta_{t-1}) \in P^1_{t-1}, \\
  \vdots \\
  x^K_{t,0} + \sum_{s=1}^{t-1} X^K_{t,s} \zeta_s, & \text{if } (\zeta_1, \ldots, \zeta_{t-1}) \in P^K_{t-1}.
\end{cases}
\]

In contrast to a single linear decision rule, finite adaptability allows for greater degrees of freedom at a greater computational cost. Indeed, for each region \( P^k \), we choose a separate linear decision rule which is locally optimal for that region. To accommodate integer decision variables, we restrict the corresponding component of each \( x^k_{t,0} \) to be integer and restrict the associated rows of each matrix \( X^k_{t,s} \) to be zero.

A complication of finite adaptability is that we may not have enough information at any intermediary stage to determine which region \( P^k \) 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 \( (\zeta_1, \ldots, \zeta_{t-1}) \), and there may be two or more regions of the partition for which their projections \( P^k_{t-1} \) and \( P^{k'}_{t-1} \) are overlapping. Fortunately, the following proposition shows that the aforementioned complication caused by overlapping projections can be resolved by adding constraints of the form \( x^k_t = x^{k'}_t \) and \( X^k_{t,s} = X^{k'}_{t,s} \) for every \( 1 \leq s < t \) when the regions \( P^k \) and \( P^{k'} \) are indistinguishable at stage \( t \).
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 \( (\zeta_1, \ldots, \zeta_{t-1}) = (\zeta'_1, \ldots, \zeta'_{t-1}) \), and \( \zeta \in \text{int}(P^k) \) or \( \zeta' \in \text{int}(P^{k'}) \) hold, then we must enforce the constraints that \( x^k_{t,0} = x^{k'}_{t,0} \) and \( x^k_{t,s} = x^{k'}_{t,s} \) for all \( 1 \leq s < t \) at stage \( t \) as the two regions cannot be distinguished with the uncertain parameters realized by that stage. Otherwise, we do not need to enforce any constraints at stage \( t \) for this pair.

For brevity, we let \( \mathcal{T}(P^1, \ldots, P^K) \) denote the collection of tuples \((k, k', t)\) for which \( P^k \) and \( P^{k'} \) cannot be distinguished at stage \( t \), which we assume can be tractably computed.

We now extend the approach of finite adaptability to Problem (2). Let \( P^1, \ldots, P^K \) be a given partition of \( \Xi \), and let the intersections between regions of the partition and uncertainty sets be denoted by \( K^j := \{ k \in [K] : U_{N}^j \cap P^k \neq \emptyset \} \). For the setting where \( c_i(\xi) \) and \( A_i(\xi) \) do not depend on the uncertain parameters, the resulting linear decision rule approximation of Problem (2) is given by

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{N} \sum_{j=1}^{N} \max_{k \in K^j} \max_{\zeta \in U_{N}^j \cap P^k} \sum_{t=1}^{T} c_t \cdot \left( x^k_{t,0} + \sum_{s=1}^{t-1} x^k_{t,s} \zeta_s \right) \\
\text{subject to} & \quad \sum_{t=1}^{T} A_i \left( x^k_{t,0} + \sum_{s=1}^{t-1} x^k_{t,s} \zeta_s \right) \leq b(\zeta) \quad \forall \zeta \in \bigcup_{j=1}^{N} U_{N}^j \cap P^k, \quad k \in [K] \\
& \quad x^k_t = x^{k'}_t, \quad x^k_{t,s} = x^{k'}_{t,s} \quad \forall (k, k', t) \in \mathcal{T}(P^1, \ldots, P^K), \quad 1 \leq s < t.
\end{align*}
\]

where the decision variables are \( x^k_{t,0} \in \mathbb{R}^{n_t} \) and \( x^k_{t,s} \in \mathbb{R}^{n_t \times d_s} \) for all \( 1 \leq s < t \) and \( k \). Extensions for general \( c_i(\xi) \) and \( A_i(\xi) \) can be found in Appendix F. Speaking intuitively, the approximation gap between Problem (2) and Problem (5) depends on the selection and granularity of the partition. By choosing partitions with a greater number of regions, Problem (5) may result in a tighter approximation of Problem (2), although this comes with an increase in problem size. To select partitions, we suggest a similar iterative heuristic to those of Postek and Hertog (2016) and Bertsimas and Dunning (2016). We present the details and theoretical justification of this heuristic in Appendix F in the context of Problem (2). The tractability and approximation quality of the proposed iterative approach is demonstrated in Section 7.

Finally, to obtain a scalable reformulation of Problem (5), we employ the same techniques as Section 5.1 to exploit the common structure of the uncertainty sets. We will assume that the intersections between the regions of the partition and uncertainty sets take a rectangular form:

**Assumption 5.** The intersection between each uncertainty set and region of the partition either has the form \( U_{N}^j \cap P^k := \{ \zeta \in \mathbb{R}^d : \ell^k \leq \zeta \leq u^k \} \) or is empty.
We remark that this assumption can be guaranteed under the same conditions as Assumption 4 when the partition’s regions are constructed as hyperrectangles. We now present the main result, which shows that Problem (5) to be reformulated as a finite-dimensional linear optimization problem which scales lightly in the number of sample paths $N$ as well as the number of regions $K$.

**Theorem 4.** If Assumption 5 holds, then (5) can be reformulated by adding at most $O(N + Kmd)$ auxiliary continuous decision variables and $O(m \sum_{j=1}^{N} |K_j| + Kmd)$ linear constraints. The reformulation is

\[
\text{minimize } \frac{1}{N} \sum_{j=1}^{N} v_j \\
\text{subject to } \sum_{t=1}^{T} (u_{i,t}^{jk} \cdot \alpha_{t}^{k} - \ell_{i,t}^{k} \cdot \beta_{i}^{k} + c_{t} \cdot x_{t,0}^{k}) \leq v_j \quad j \in [N], \ k \in K_j
\]

\[
\alpha_{t}^{k} - \beta_{t}^{k} = \sum_{s=t+1}^{T} (X_{s,t}^{k})^\top c_s \quad t \in [T], \ k \in [K]
\]

\[
\sum_{t=1}^{T} (M_{i,t}^{k} u_{i,t}^{k} - \Lambda_{i,t}^{k} \ell_{i,t}^{k} + A_{i,t} x_{t,0}^{k}) \leq b^0 \quad j \in [N], \ k \in K_j
\]

\[
M_{i,t}^{k} - \Lambda_{i,t}^{k} = -B_t + \sum_{s=t+1}^{T} A_s X_{t,s}^{k} \quad t \in [T], \ k \in [K]
\]

\[
x_t^k = x_{t'}^k, \quad X_{t,s}^k = X_{t',s} \quad (k, k', t) \in \mathcal{T}(P_1, \ldots, P_K), \ 1 \leq s < t.
\]

where the auxiliary decision variables are $v \in \mathbb{R}^N$ as well as $\alpha^k := (\alpha_1^k, \ldots, \alpha_T^k), \beta^k := (\beta_1^k, \ldots, \beta_T^k) \in \mathbb{R}_+^d$ and $M^k := (M_1^k, \ldots, M_T^k), \Lambda^k := (\Lambda_1^k, \ldots, \Lambda_T^k) \in \mathbb{R}_+^{m \times d}$ for each $k \in [K]$. Note that $b(\zeta) := b^0 + \sum_{t=1}^{T} B_t \zeta_t \in \mathbb{R}^m$.

**Proof.** The proof follows from similar reasoning as Theorem 3 and is thus omitted.

This result suggests that Problem (2) with finite adaptability is highly scalable, in the sense that the size of the resulting reformulation scales lightly in the number of sample paths $N$. Under the same assumptions of Theorem 4, we remark that $\ell^i, u^i$, and $\mathcal{T}(P_1, \ldots, P_K)$ can be computed efficiently by computing the intersection of each uncertainty set and region of the partition.

6. Relationships with Distributionally Robust Optimization

In the previous sections, we discussed the theoretical underpinnings and computational tractability of Problem (2) as a data-driven approach to multi-stage stochastic linear optimization. An attractive aspect of the proposed approach is its simplicity, interpretable as a straightforward robustification of historical sample paths. In this section, we explore connections between our data-driven approach to multi-stage stochastic linear optimization and implications of our results in the context of distributionally robust optimization.
Our exposition focuses on the following formulation of multi-stage distributionally robust linear optimization:

$$\begin{align*}
\text{minimize} & \quad \sup_{\mathbf{Q} \in \mathcal{A}} \mathbb{E}_\mathbf{Q} \left[ \sum_{t=1}^{T} c_t(\xi) \cdot x_t(\xi_1, \ldots, \xi_{t-1}) \right] \\
\text{subject to} & \quad \sum_{t=1}^{T} A_t(\xi)x_t(\xi_1, \ldots, \xi_{t-1}) \leq b(\xi) \quad \mathbb{Q}\text{-a.s.}, \forall \mathbf{Q} \in \mathcal{A}.
\end{align*}$$

(6)

Intuitively speaking, this framework chooses the decision rules which minimize the expected cost with respect to an adversarially chosen probability distribution from an ambiguity set. The requirement that the constraints hold almost surely for every distribution in the ambiguity set ensures that the objective function will evaluate the cost function on realizations of the stochastic process where the decision rules are feasible. Examples of this formulation in multi-stage and data-driven two-stage problems include Bertsimas et al. (2018c) and Hanasusanto and Kuhn (2018).

Our following discussion focuses on ambiguity sets which are constructed using historical data and Wasserstein-based distances between probability distributions. Given two bounded probability distributions, their $\infty$-Wasserstein distance is defined as

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

where the essential supremum of the joint distribution is given by

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

For any $p \in [1, \infty)$, the $p$-Wasserstein distance between two probability distributions is defined as

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

with marginals $\mathbf{Q}$ and $\mathbf{Q}'$, respectively.

For technical details on these distances, we refer the reader to Givens and Shortt (1984). For any $p \in [1, \infty]$, let the $p$-Wasserstein ambiguity set be defined as

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

where $\epsilon_N \geq 0$ is a robustness parameter which controls the size of the ambiguity set and $\hat{\mathbf{P}}_N$ is the empirical probability distribution which assigns equal weight to each of the historical sample paths $\hat{\xi}^1, \ldots, \hat{\xi}^N$. We henceforth refer to Problem (6) with the $p$-Wasserstein ambiguity set as $p$-WDRO.

As discussed at the end of Section 1.2, there are relatively few previous convergence guarantees for distributionally robust optimization with the $\infty$-Wasserstein ambiguity set, even for single-stage
problems. Indeed, when the underlying distribution is unbounded, the $\infty$-Wasserstein ambiguity set will never contain the true distribution, even as $N$ tends to infinity, since the distance $d_\infty(P, \hat{P}_N)$ from the true to the empirical distribution will always be infinite. Thus, except under stronger assumptions than Assumption 1, the techniques used by Esfahani and Kuhn (2018, Theorems 3.5 and 3.6) to establish finite-sample and convergence guarantees for the 1-Wasserstein ambiguity set do not appear to extend to the $\infty$-Wasserstein ambiguity set. Nonetheless, distributionally robust optimization with the $\infty$-Wasserstein ambiguity set has recently received interest in the context of regularization and adversarial training in machine learning (Gao et al. 2017, Staib and Jegelka 2017).

The following main result of this section shows that Problem (2), under a particular construction of uncertainty sets, can also be interpreted as Problem (6) with the $\infty$-Wasserstein ambiguity set.

**Theorem 5.** Problem (2) with uncertainty sets of the form

$$U_j^i := \{ \zeta \equiv (\zeta_1, \ldots, \zeta_T) \in \Xi : \| \zeta - \hat{\zeta}^j \| \leq \epsilon_N \}$$

is equivalent to $\infty$-WDRO.

**Proof.** See Appendix G. □

Therefore, as a byproduct of Theorem 2 from Section 4, we have obtained general convergence guarantees for distributionally robust optimization using the $\infty$-Wasserstein ambiguity set under mild probabilistic assumptions. We refer the reader to our companion paper Bertsimas et al. (2019) for additional performance guarantees under stronger probabilistic assumptions.

For comparison, we now show that similar asymptotic optimality guarantees for multi-stage stochastic linear optimization are not obtained by $p$-WDRO for any $p \in [1, \infty)$. Indeed, the following theorem shows that the constraints induced by such an approach are overly conservative in general.

**Theorem 6.** If $p \in [1, \infty)$ and $\epsilon_N > 0$, then a decision rule is feasible for $p$-WDRO only if

$$\sum_{i=1}^{T} A_i(\zeta)x_i(\zeta_1, \ldots, \zeta_{i-1}) \leq b(\zeta) \quad \forall \zeta \in \Xi.$$

**Proof.** See Appendix H. □

As discussed in Section 2, the set $\Xi$ is not necessarily a tight approximation of the true (unknown) support of the stochastic process, and may be strictly and significantly larger. Thus, the constraints induced from $p$-WDRO with $p \in [1, \infty)$ may eliminate optimal or high-quality decision rules for Problem (1). Consequently, $p$-WDRO with $p \in [1, \infty)$ is not asymptotically optimal for multi-stage stochastic linear optimization in general. We conclude this section with two further remarks.
Remark 1. If we relaxed the constraints of $p$-WDRO in an attempt to decrease its conservatism, then the resulting decision rules are not guaranteed to be feasible for the stochastic problem. Thus, the finite-sample guarantees provided by Esfahani and Kuhn (2018, Equation 2), which served as one of the principle justifications for using $p$-WDRO, would no longer provide meaningful insight into the true out-of-sample performance of this decision rule. □

Remark 2. The conservatism of $p$-WDRO can also lead to suboptimal decisions even for problems where uncertainty does not impact feasibility. Indeed, consider the problem

\[
\begin{align*}
\text{minimize} & \quad \mathbb{E} [x_2(\xi_1) + 2x_3(\xi_1, \xi_2)] \\
\text{subject to} & \quad x_2(\xi_1) + x_3(\xi_1, \xi_2) \geq \xi_1 + \xi_2 \quad \text{a.s.} \\
& \quad x_2(\xi_1) + x_3(\xi_1, \xi_2) \geq \xi_1 - \xi_2 \quad \text{a.s.}
\end{align*}
\]

We observe that $x_2(\xi_1) = \xi_1$ and $x_3(\xi_1, \xi_2) = |\xi_2|$ are feasible decision rules, regardless of the underlying probability distribution. Suppose that the probability distribution and support of $(\xi_1, \xi_2)$ is unknown, and our only information comes from historical data. If we approximate this stochastic problem using $p$-WDRO for any $p \in [1, \infty)$ and linear decision rules, we are tasked with solving

\[
\begin{align*}
\text{minimize} & \quad \sup_{\zeta \in \mathcal{A}_N} \mathbb{E}_Q \left[ (x_{2,0} + x_{2,1}\zeta_1) + 2(x_{3,0} + x_{3,1}\zeta_1 + x_{3,2}\zeta_2) \right] \\
\text{subject to} & \quad (x_{2,0} + x_{2,1}\zeta_1) + (x_{3,0} + x_{3,1}\zeta_1 + x_{3,2}\zeta_2) \geq \zeta_1 + \zeta_2 \quad \forall \zeta \in \mathbb{R}^2 \\
& \quad (x_{2,0} + x_{2,1}\zeta_1) + (x_{3,0} + x_{3,1}\zeta_1 + x_{3,2}\zeta_2) \geq \zeta_1 - \zeta_2 \quad \forall \zeta \in \mathbb{R}^2.
\end{align*}
\]

It follows from identical reasoning as Bertsimas et al. (2018c, Section 3) that there are no linear decision rules which are feasible for the above optimization problem. In particular, the above optimization problem will remain infeasible even if the true support of the random variable happens to be bounded but the bound is unknown. In contrast, the sample robust optimization approach (Problem (2)) to this example will always have a feasible linear decision rule. □

7. Applications and Numerical Experiments

In this section, we demonstrate the practical value of the proposed data-driven approach for addressing real-world multi-stage stochastic linear optimization problems. Specifically, we assess the average out-of-sample performance and computational tractability of proposed data-driven approach in applications where uncertainty is correlated across time and our only knowledge comes from historical data. In Section 7.1, we consider a multi-stage stochastic inventory problem with (unknown) autoregressive demands. In Section 7.2, we consider a computationally demanding multi-stage mixed-integer stochastic lot sizing problem for new short lifecycle products.

We compare the following data-driven approaches for multi-stage stochastic linear optimization:
• *(SRO)* The proposed data-driven approach (Problem (2)) for solving multi-stage stochastic linear optimization. The uncertainty sets are constructed as described in Section 3 where the robustness parameter $\epsilon_N$ is chosen by cross-validation (in Section 7.1) and by Assumption 2 (in Section 7.2). We solve this approach using linear decision rules (in Section 7.1) and finite adaptability (in Section 7.2). The partitions in finite adaptability are constructed using the heuristic iterative algorithm from Appendix F with a single iteration of partitioning.

• *(Approx PCM)* A data-driven variation on the approach of Bertsimas et al. (2018c). In this approach, we obtain an approximation of multi-stage stochastic linear optimization by solving Problem (6) from Section 6 with ambiguity sets constructed as the joint distributions with the same first and second moments (mean and covariance) as those estimated from the historical data. We solve this approach using lifted linear decision rules, as described in Bertsimas et al. (2018c, Section 3).

• *(RDDP)* The robust data-driven dynamic programming approach proposed by Hanasusanto and Kuhn (2013). This approach estimates the cost-to-go functions by applying kernel regression to the historical sample paths. We solve the resulting dynamic program using value function approximation, as described in Hanasusanto and Kuhn (2013, Section 4). Since RDDP requires both input sample paths and initial state paths, we use half of the training dataset as the input sample paths, and the other half to generate the state paths with lifted linear decision rules obtained by Approx PCM. The approach also requires a robustness parameter $\gamma$, which we choose to be either $\gamma = 0$ (DDP) or $\gamma = 10$ (RDDP).

• *(WDRO)* The 1-WDRO approach described in Section 6. In this approach, we obtain an approximation of multi-stage stochastic linear optimization by solving Problem (6) from Section 6 where the ambiguity sets are constructed as the 1-Wasserstein ambiguity set with the $\ell_1$ norm. We solve the resulting problem using the finite adaptability approach described in Section 5 with a reformulation described in Appendix I. We use the same robustness parameter and partitions as SRO.

In each experiment, the above approaches are applied to the same training datasets and their produced decision rules are evaluated on a common testing dataset. Moreover, in each experiment, all training and testing sample paths are drawn independently from the joint probability distribution of the multi-stage stochastic linear optimization problem. Note that decision rules produced by SRO are evaluated only on sample paths in the testing dataset which result in feasible decisions. We report on the percentage of sample paths in the testing dataset which are feasible for these decision rules in each experiment. All algorithms were implemented in JuMP (Lubin and Dunning 2015) and solved using Gurobi 8.1.1 (Gurobi Optimization 2018).
7.1. Multi-stage stochastic inventory management

We first consider a multi-stage stochastic inventory management problem with autoregressive demand models from See and Sim (2010). The purpose of this section is to illustrate the optimality and tractability of the proposed data-driven approach with linear decision rules.

**Description.** We are tasked with managing the inventory of a product with uncertain demand over a finite planning horizon of $T$ time periods. At the beginning of the $t$-th time period, let $I_t$ denote the number of product units in inventory. We then select a production quantity of $x_t \in [0, \bar{x}_t]$ units with zero lead time at a cost of $c_t$ per unit. The product demand $\xi_t \geq 0$ is then revealed, and the inventory in the following period is updated to $I_t + 1 = I_t + x_t - \xi_t$. In addition to production costs, we incur in each time period a holding cost of $h_t$ per unit and a backorder cost of $b_t$ per unit. We begin initially with zero units of inventory. Our goal is to minimize the expected cost over the planning horizon, given by the following multi-stage stochastic linear optimization problem:

$$
\minimize_{x, I, y} \mathbb{E} \left[ \sum_{t=1}^{T} (c_t x_t(\xi_1, \ldots, \xi_{t-1}) + y_{t+1}(\xi_1, \ldots, \xi_t)) \right]
$$

subject to

$$
\begin{align*}
I_{t+1}(\xi_1, \ldots, \xi_t) &= I_t(\xi_1, \ldots, \xi_{t-1}) + x_t(\xi_1, \ldots, \xi_{t-1}) - \xi_t \quad \text{a.s., } \forall t \in [T] \\
y_{t+1}(\xi_1, \ldots, \xi_t) &\geq h_t I_{t+1}(\xi_1, \ldots, \xi_t) \quad \text{a.s., } \forall t \in [T] \\
y_{t+1}(\xi_1, \ldots, \xi_t) &\geq -b_t I_{t+1}(\xi_1, \ldots, \xi_t) \quad \text{a.s., } \forall t \in [T] \\
0 &\leq x_t(\xi_1, \ldots, \xi_{t-1}) \leq \bar{x}_t \quad \text{a.s., } \forall t \in [T],
\end{align*}
$$

(7)

Note that we have introduced epigraph decision rules $y_{t+1} : \mathbb{R}^t \to \mathbb{R}$ to capture the holding and backorder costs from period $t$. In our experiments, we eliminate the inventory decision rules by recursively applying the equality constraints.

**Experiments.** We preform computational simulations using the same parameters and data generation as See and Sim (2010). The demand follows a nonstationary autoregressive 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 $[-\bar{\varsigma}, \bar{\varsigma}]$. The parameters of the stochastic process are $\mu = 200$ and $\bar{\varsigma} = 20$ (when $T = 5$) and $\mu = 200$ and $\bar{\varsigma} = 10$ (when $T = 10$). The capacities and costs are $\bar{x}_t = 260$, $c_t = 0.1$, $h_{t+1} = 0.02$ for all $t \in [T]$, $b_t = 0.2$ for all $t \in [T-1]$, and $b_{T+1} = 2$. We use $\Xi = \mathbb{R}^T_+$ for all approaches. For SRO and WDRO, we use linear decision rules and choose the robustness parameter $\epsilon_N \geq 0$ using leave-one-out cross-validation. Specifically, for each training dataset, we compute the distance from each historical sample path to the closest among other historical sample paths, and then take the maximum of those distances as the value of $\epsilon_N$.

**Results.** In Table 1 and Figure 1, we present the average out-of-sample costs and computation times by the various data-driven approaches. The results show that SRO achieves the best average...
Table 1 Multi-stage stochastic inventory problem: average out-of-sample cost.

<table>
<thead>
<tr>
<th>$T$</th>
<th>$\alpha$</th>
<th>Approach</th>
<th>Size of training dataset (N)</th>
<th></th>
<th></th>
<th></th>
<th>DP</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0</td>
<td>SRO</td>
<td>109.2(1.5)</td>
<td>108.4(0.9)</td>
<td>107.6(0.6)</td>
<td>107.1(0.4)</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approx PCM</td>
<td>109.2(1.1)</td>
<td>108.7(0.5)</td>
<td>108.5(0.3)</td>
<td>108.5(0.3)</td>
<td></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>
<td></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>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td></td>
<td>SRO</td>
<td>108.9(1.3)</td>
<td>108.6(0.8)</td>
<td>108.5(0.7)</td>
<td>108.3(0.6)</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approx PCM</td>
<td>113.3(1.9)</td>
<td>112.5(0.7)</td>
<td>112.3(0.5)</td>
<td>112.2(0.4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DDP</td>
<td>2278.0(409.7)</td>
<td>1297.6(918.9)</td>
<td>617.7(637.2)</td>
<td>194.9(217.8)</td>
<td></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>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td>SRO</td>
<td>109.7(1.7)</td>
<td>110.1(1.4)</td>
<td>110.1(1.2)</td>
<td>110.1(1.1)</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approx PCM</td>
<td>117.8(2.5)</td>
<td>116.8(1.0)</td>
<td>116.5(0.7)</td>
<td>116.4(0.6)</td>
<td></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>
<td></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>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>SRO</td>
<td>212.4(1.5)</td>
<td>208.5(0.9)</td>
<td>207.5(0.7)</td>
<td>206.8(0.3)</td>
<td>206</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approx PCM</td>
<td>208.0(1.0)</td>
<td>207.5(0.4)</td>
<td>207.3(0.3)</td>
<td>207.3(0.2)</td>
<td></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(263.5)</td>
<td></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(266.3)</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td></td>
<td>SRO</td>
<td>208.7(1.2)</td>
<td>208.3(0.8)</td>
<td>208.1(0.6)</td>
<td>208.1(0.5)</td>
<td>206</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approx PCM</td>
<td>214.4(2.2)</td>
<td>213.4(0.8)</td>
<td>213.0(0.6)</td>
<td>212.8(0.5)</td>
<td></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>
<td></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>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td></td>
<td>SRO</td>
<td>210.1(1.8)</td>
<td>210.1(1.4)</td>
<td>210.1(1.1)</td>
<td>210.2(1.0)</td>
<td>207</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Approx PCM</td>
<td>222.2(3.4)</td>
<td>220.5(1.3)</td>
<td>220.0(1.0)</td>
<td>219.7(0.7)</td>
<td></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>
<td></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>
<td></td>
</tr>
</tbody>
</table>

Mean (standard deviation) for the average out-of-sample cost for the multi-stage stochastic inventory problem over 100 training datasets. The average out-of-sample cost for each training dataset is computed on a test set of 10,000 sample paths. Note that the out-of-sample cost was computed directly from the ordering policies $x_1, \ldots, x_T$ produced by each approach. The column DP presents the dynamic programming approximation of the optimal cost of Problem (7) from See and Sim (2010, Appendix EC.1) with an accuracy of ±1%.

out-of-sample cost and computation time in each instance. In particular, the average out-of-sample cost of the linear decision rules from SRO nearly matches the optimal cost of Problem (1). Table 2 shows that decision rules produced by SRO are also feasible for most or all of the sample paths in the testing dataset. Note that the results exclude WDRO with linear decision rules, as this approach resulted in an infeasible optimization problem in every instance.

7.2. Multi-stage stochastic lot sizing

We next consider a computationally demanding mixed-integer stochastic lot sizing problem for short lifecycle products. The purpose of this section is to illustrate the computational tractability and optimality of the proposed data-driven approach in a problem with continuous and integer decision variables and where future uncertainty can affect feasibility.
Figure 1 Multi-stage stochastic inventory management: computation times.

Note. Computation times for data-driven approaches to the multi-stage stochastic inventory management problem with ten stages ($T = 10$) and no correlation between time steps ($\alpha = 0$). Similar computation times were observed for other choices of $\alpha$. The graph shows the mean value of the computation times over 100 training datasets for each value of $N$.

Table 2 Multi-stage stochastic inventory management: out-of-sample feasibility.

<table>
<thead>
<tr>
<th>$T$</th>
<th>$\alpha$</th>
<th>Size of training dataset (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>99.91(0.49)</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>99.59(1.35)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>98.71(2.80)</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>100.00(0.00)</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>99.79(0.47)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>98.77(2.24)</td>
</tr>
</tbody>
</table>

Mean (standard deviation) of the percentage of the 10,000 sample paths in the testing dataset which were feasible for the ordering policy generated by SRO. The mean and standard deviation are computed over 100 training datasets for each value of $N$, $T$, $\alpha$.

Description. We consider a manufacturer which produces products with short lifecycles (e.g., seasonal fashion apparel or computer hardware) for a retailer. While the demand trajectory from the supplier of a new product is unknown, the manufacturer is endowed with historical sales over the lifecycle of similar products, visualized in Figure 2. The goal is to determine the manufacturer’s production decisions over a product’s lifecycle to satisfy the retailer’s orders while minimizing production and holding costs. More formally, let the product lifecycle be discretized into $T$ periods. In each period $t \in \{1, \ldots, T\}$, we (the manufacturer) start with $I_1 \geq 0$ units of product in inventory ($I_1 = 0$) and observe the demand of $\xi_t \geq 0$ units from the retailer. We are obliged to satisfy all of the demand, and any excess inventory in each period incurs a holding cost of $h$ per unit. To satisfy the demand, we have several options. We may begin production of $x_{t+1} \geq 0$ units at a cost of $c$ per unit,
which will become available in the following period. Furthermore, for each \( i \in \{1, \ldots, Q\} \), we may produce a lot of \( q_i \geq 0 \) units at a cost of \( p_i \geq 0 \) per unit, which will become available immediately. We denote the binary decisions for whether to produce each lot by \( z_{t+1,i} \), \( 1 \leq i \leq Q \). The resulting multi-stage stochastic linear optimization problem is given by

\[
\minimize_{\mathbf{x}, \mathbf{I}, \mathbf{z}} \quad \mathbb{E} \left[ \sum_{t=1}^{T} (c x_t(\xi_{1:t-1}) + h I_{t+1}(\xi_{1:t}) + \sum_{i=1}^{Q} p_i q_i z_{t+1,i}(\xi_{1:t})) \right]
\]

subject to

\[
I_{t+1}(\xi_{1:t}) = I_t(\xi_{1:t-1}) + x_t(\xi_{1:t-1}) + \sum_{i=1}^{Q} q_i z_{t+1,i}(\xi_{1:t}) - \xi_t \quad \text{a.s., } \forall t \in [T]
\]

\[
I_{t+1}(\xi_{1:t}), x_t(\xi_{1:t-1}) \geq 0, \ z_{t+1,i}(\xi_{1:t}), \ldots, z_{t+1,Q}(\xi_{1:t}) \in \{0,1\}, \ \text{a.s., } \forall t \in [T],
\]

where we use the shorthand notation \( \xi_{1:t} := (\xi_1, \ldots, \xi_t) \). In our experiments, we eliminate the inventory decision rules by recursively applying the equality constraints.

**Experiments.** We construct the demand process as a mixture of three stochastic processes (“low”, “medium”, and “high”). More precisely, each sample path of the stochastic process is drawn from one of the three joint distributions, each with equal probability of selection. The three stochastic processes are defined as

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

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

where \( \epsilon_1, \ldots, \epsilon_T \) denote independently and identically distributed Gaussian random variables with mean zero and standard deviation of 7 units. To ensure that Problem (8) is feasible, we use rejection sampling to restrict the demands in each period to be upper bounded by 150 units. The remaining
Note. Average out-of-sample cost and computation time for the multi-stage stochastic lot sizing problem with $T = 8$. The left graph shows the average out-of-sample cost on test set of 10,000 sample paths. The bottom graph shows the average solver time in seconds. The robustness parameters for SRO and WDRO are computed as $\epsilon_N = \kappa N^{-\frac{1}{2}}$ for $\kappa = 50$. All results for each choice of $N$ are averaged over 100 training datasets. 

Results. In Figure 3, we present the average out-of-sample cost and computation times for the various approaches with $T = 8$. The results demonstrate that SRO produces decision rules with an average cost that significantly outperforms the alternative approaches. In particular, the average out-of-sample cost of these decision rules improves significantly with the size of the training dataset ($N$). We observe that the computation time for SRO grows linearly in the size of the training dataset for $N \geq 100$. In Table 3, we demonstrate that, under our choice of robustness parameter, the decision rules produced by SRO are feasible with high probability.

8. Conclusion

In this work, we presented a general-purpose data-driven approach for solving multi-stage stochastic linear optimization problems, including those with integer decisions, where uncertainty is corre-
Table 3  Multi-stage stochastic lot sizing: out-of-sample feasibility.

<table>
<thead>
<tr>
<th>Size of training dataset (N)</th>
<th>Feasibility percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>99.92(0.14)</td>
</tr>
<tr>
<td>50</td>
<td>99.97(0.06)</td>
</tr>
<tr>
<td>75</td>
<td>99.97(0.04)</td>
</tr>
<tr>
<td>100</td>
<td>99.98(0.03)</td>
</tr>
<tr>
<td>125</td>
<td>99.98(0.03)</td>
</tr>
<tr>
<td>150</td>
<td>99.99(0.02)</td>
</tr>
<tr>
<td>175</td>
<td>99.99(0.02)</td>
</tr>
<tr>
<td>200</td>
<td>99.99(0.01)</td>
</tr>
</tbody>
</table>

Mean (standard deviation) of the percentage of the 10,000 sample paths in the testing dataset which were feasible for the decision rules generated by SRO. The mean and standard deviation are computed over 100 training datasets for each value of $N$.

lated across time. For this approach, we established nonparametric convergence guarantees and scalable approximation algorithms. The practical value of the proposed approach was illustrated by computational examples motivated by real-world applications, demonstrating that the proposed data-driven approach can produce high-quality decisions in reasonable computation times. Apart from these contributions, the paper also presented general guarantees for distributionally robust optimization using the $\infty$-Wasserstein ambiguity set, which may have implications in areas such as statistical learning theory.

Intriguingly, the proposed data-driven framework proposed in this paper also provides a new approach for approximating multi-stage stochastic linear optimization problems. Namely, suppose that the underlying distribution of the stochastic process was known. Then, this paper shows that one can approximate the corresponding multi-stage stochastic linear optimization problem by (i) obtaining sample paths from the stochastic process by Monte Carlo simulation, and (ii) approximately solving Problem (2). This approach thus offers an alternative to the traditional approximation approach of constructing scenario trees, and provides a new and exciting opportunity for leveraging advances in robust optimization to solve multi-stage stochastic linear optimization.

References


Appendix A: Examples of Assumption 3 from Section 4.1

In this appendix, we show that every multi-stage stochastic linear optimization problem considered in this paper satisfies Assumption 3.

A.1. Example 1 from Section 3

Consider the sample robust optimization problem

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{N} \sum_{j=1}^{N} \sup_{\zeta \in U_j} \{ x_2(\zeta_1) + 2x_3(\zeta_1, \zeta_2) \} \\
\text{subject to} & \quad x_2(\zeta_1) + x_3(\zeta_1, \zeta_2) \geq \zeta_1 + \zeta_2 \quad \forall \zeta \in \bigcup_{j=1}^{N} U_j \\
& \quad x_2(\zeta_1), x_3(\zeta_1, \zeta_2) \geq 0 \quad \forall \zeta \in \bigcup_{j=1}^{N} U_j.
\end{align*}
\]

We observe that the decisions must be nonnegative for every realization in the uncertainty sets. Moreover, the following constraints can be added to the above problem without affecting its optimal cost:

\[
\begin{align*}
x_2(\zeta_1) & \leq \sup_{\zeta' \in \bigcup_{j=1}^{N} U_j} \{ \zeta'_1 + \zeta'_2 \} \quad \forall \zeta \in \bigcup_{j=1}^{N} U_j, \\
x_3(\zeta_1, \zeta_2) & \leq \sup_{\zeta' \in \bigcup_{j=1}^{N} U_j} \{ \zeta'_1 + \zeta'_2 \} \quad \forall \zeta \in \bigcup_{j=1}^{N} U_j.
\end{align*}
\]

Indeed, the above constraints ensure that we are never purchasing inventory which exceeds the maximal \(\zeta_1 + \zeta_2\) which can be realized in the uncertainty sets. Thus, we have shown that Assumption 3 holds.

A.2. Example 2 from Section 4.3

Consider the sample robust optimization problem

\[
\begin{align*}
\text{minimize} & \quad x_1 \\
\text{subject to} & \quad x_1 \geq \zeta_1 \quad \forall \zeta \in \bigcup_{j=1}^{N} U_j.
\end{align*}
\]

We observe that an optimal solution to this problem is \(x_1 = \max_{\zeta_1 \in \bigcup_{j=1}^{N} U_j} \lceil \zeta_1 \rceil\), and thus the constraint

\[
x_1 \leq \max_{\zeta_1 \in \bigcup_{j=1}^{N} U_j} \lceil \zeta_1 \rceil + 1
\]

can be added to the above problem without affecting its objective value. We conclude that Assumption 3 holds.

A.3. Inventory example from Section 7.1

Consider the sample robust optimization problem

\[
\begin{align*}
\text{minimize} & \quad x_1 \\
\text{subject to} & \quad x_1 \geq \zeta_1 \quad \forall \zeta \in \bigcup_{j=1}^{N} U_j.
\end{align*}
\]
Indeed, the above constraints ensure that we are never purchasing inventory in a given stage which exceeds further imply that

\[
I_{t+1}(\zeta_1, \ldots, \zeta_t) = I_t(\zeta_1, \ldots, \zeta_{t-1}) + x_t(\zeta_1, \ldots, \zeta_{t-1}) - \zeta_t \quad \forall \zeta \in \cup_{j=1}^N U_j, \forall t \in [T]
\]

\[
y_{t+1}(\zeta_1, \ldots, \zeta_t) \geq h_{t+1}(\zeta_1, \ldots, \zeta_t) \quad \forall \zeta \in \cup_{j=1}^N U_j, \forall t \in [T]
\]

\[
y_{t+1}(\zeta_1, \ldots, \zeta_t) \geq -b_{t+1}(\zeta_1, \ldots, \zeta_t) \quad \forall \zeta \in \cup_{j=1}^N U_j, \forall t \in [T]
\]

\[
0 \leq x_t(\zeta_1, \ldots, \zeta_{t-1}) \leq \bar{x}_t \quad \forall \zeta \in \cup_{j=1}^N U_j, \forall t \in [T].
\]

For any feasible decision rule to the above problem and for each stage \(t\), we observe that the following constraint is satisfied:

\[
- \sup_{\zeta' \in \cup_{j=1}^N U_j} \sum_{s=1}^T \zeta'_s \leq I_{t+1}(\zeta_1, \ldots, \zeta_t) \leq \sum_{s=1}^T \bar{x}_s \quad \forall \zeta \in \cup_{j=1}^N U_j.
\]

Indeed, the lower bound follows from the case where \(x_t(\zeta_1, \ldots, \zeta_{t-1}) = 0\) for every realization in the uncertainty sets and the maximum demand in the uncertainty sets is observed. Moreover, we can without loss of generality impose the constraint that

\[
0 \leq y_{t+1}(\zeta_1, \ldots, \zeta_t) = \max \{ h_{t+1}(\zeta_1, \ldots, \zeta_t), -b_{t+1}(\zeta_1, \ldots, \zeta_t) \} \quad \forall \zeta \in \cup_{j=1}^N U_j.
\]

Applying the aforementioned bounds on \(I_{t+1}(\zeta_1, \ldots, \zeta_t)\) over the uncertainty sets, we conclude that Assumption 3 holds.

### A.4. Lot-sizing example from Section 7.2

Consider the sample robust optimization problem

\[
\text{minimize} \quad \frac{1}{N} \sum_{j=1}^N \sup_{\zeta \in U_j} \sum_{t=1}^T \left( c_{x,t}(\zeta_{t-1}) + h_{t+1}(\zeta_{t}) + \sum_{i=1}^Q p_i q_i z_{t+1,i}(\zeta_{t}) \right)
\]

subject to

\[
I_{t+1}(\zeta_{1:t}) = I_t(\zeta_{1:t-1}) + x_t(\zeta_{1:t-1}) + \sum_{i=1}^Q p_i q_i z_{t+1,i}(\zeta_{t}) - \zeta_t \quad \forall \zeta \in \cup_{j=1}^N U_j, \forall t \in [T]
\]

\[
x_t(\zeta_{1:t-1}) \geq 0, \quad z_{t+1,1}(\zeta_{1:t}), \ldots, z_{t+1,Q}(\zeta_{1:t}) \in \{0,1\} \quad \forall \zeta \in \cup_{j=1}^N U_j, \forall t \in [T]
\]

We observe that the decisions must be nonnegative for every realization in the uncertainty sets. Moreover, the following constraints can be added to the above problem for each stage \(t\) without affecting its optimal cost:

\[
x_t(\zeta_1, \ldots, \zeta_{t-1}) \leq \sup_{\zeta' \in \cup_{j=1}^N U_j} \{ \zeta'_1 + \cdots + \zeta'_T \} \quad \forall \zeta \in \cup_{j=1}^N U_j.
\]

Indeed, the above constraints ensure that we are never purchasing inventory in a given stage which exceeds the maximal \(\zeta_1 + \cdots + \zeta_T\) which can be realized in the uncertainty sets. Naturally, the above constraints will further imply that

\[
I_t(\zeta_1, \ldots, \zeta_{t-1}) \leq T \left( \sup_{\zeta' \in \cup_{j=1}^N U_j} \{ \zeta'_1 + \cdots + \zeta'_T \} + \sum_{i=1}^Q q_i \right) \quad \forall \zeta \in \cup_{j=1}^N U_j, \forall t \in [T].
\]
Since the remaining decision rules map to binary values, we have shown that Assumption 3 holds.

**Appendix B: Proof of Theorem 1 from Section 4.2**

In this appendix, we present the proof of Theorem 1. Our proof techniques follow similar reasoning to Devroye and Wise (1980) and Bafilo et al. (2000) for $S_N := \bigcup_{j=1}^N U_N$, which we adapt to Assumption 1. We remark that the following theorem also provides an intermediary step for Theorem 2 from Section 4.2, the proof of which is found in Appendix D.

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

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

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

$$
A_N := \{ \zeta \in \mathbb{R}^d : \|\zeta\| \leq (\log N)^{a+1/2} \}.
$$

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

$$
R_N \mathbb{P}(\xi \not\in A_N) = R_N \mathbb{P}\left(\|\xi\| > (\log N)^{a+1/2}\right) = R_N \mathbb{P}\left(\exp(\|\xi\|^a) > \exp((\log N)^{a+1})\right) \leq \frac{bR_N}{\exp((\log N)^{a+1})} \leq \eta.
$$

The first inequality follows from Markov’s inequality and the second inequality holds for all sufficiently large $N \in \mathbb{N}$ since $a > 1$.

Next, define

$$
\alpha_N := \frac{\eta}{(\log N)^{a+1/2} \phi R_N} \quad \text{and} \quad B_N := \{ \zeta \in \mathbb{R}^d : \mathbb{P}(\|\xi - \zeta\| \leq \epsilon_N) > \alpha_N \epsilon_N^d \},
$$

where $\phi > 0$ is a constant which depends only on $d$ and will be defined shortly. We now show that $R_N \mathbb{P}(\xi \not\in B_N) \leq 2\eta$ for all sufficiently large $N \in \mathbb{N}$. Indeed, for all sufficiently large $N \in \mathbb{N},$

$$
R_N \mathbb{P}(\xi \not\in B_N) = R_N \mathbb{P}(\xi \in A_N, \xi \not\in B_N) + R_N \mathbb{P}(\xi \not\in A_N, \xi \not\in B_N)
\leq R_N \mathbb{P}(\xi \in A_N, \xi \not\in B_N) + R_N \mathbb{P}(\xi \not\in A_N)
\leq R_N \mathbb{P}(\xi \in A_N, \xi \not\in B_N) + \eta, \quad (\text{EC.1})
$$

where the final inequality follows because $R_N \mathbb{P}(\xi \not\in A_N) \leq \eta$ for all sufficiently large $N \in \mathbb{N}$. Now, choose points $\zeta^1, \ldots, \zeta^K \in A_N$ such that $\min_{j \in [K_N]} \|\zeta - \zeta^j\| \leq \epsilon_N$ 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

$$
K_N \leq \phi \left( \frac{(\log N)^{a+1/2}}{\epsilon_N} \right)^d, \quad (\text{EC.2})
$$
where $\phi > 0$ is a constant that depends only on $d$. Then, continuing from (EC.1),

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

(E.3)

where the second inequality follows from the union bound. For each $j \in [K_N]$, we have two cases to consider. First, suppose there exists a realization $\zeta \notin B_N$ such that $\|\zeta - \zeta_j\| \leq \frac{\epsilon_N}{2}$. Then,

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

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

$$\mathbb{P}\left(\|\xi - \zeta_j\| \leq \frac{\epsilon_N}{2}, \xi \notin B_N\right) = 0,$$

In each of the two cases, we have shown that

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

(E.4)

for each $j \in [K_N]$. Therefore, we combine (EC.3) and (EC.4) to obtain the following upper bound on $R_N \mathbb{P}(\xi \notin B_N)$ for all sufficiently large $N \in \mathbb{N}$:

$$R_N \mathbb{P}(\xi \notin B_N) \leq R_N K_N \alpha_N \epsilon_N^d + \eta \leq (\log N) \frac{\alpha(a+1)}{2\alpha} \phi R_N \alpha_N + \eta \leq 2\eta.$$  

(E.5)

The first inequality follows from (EC.3) and (EC.4), the second inequality follows from (EC.2), and the third inequality follows for all sufficiently large $N \in \mathbb{N}$ by the definition of $\alpha_N$.

We now prove the main result. Indeed, for all sufficiently large $N \in \mathbb{N}$,

$$R_N \mathbb{P}(\xi \notin S_N) = R_N \mathbb{P}(\xi \notin S_N, \xi \in B_N) + R_N \mathbb{P}(\xi \notin S_N, \xi \notin B_N) \leq R_N \mathbb{P}(\xi \notin S_N, \xi \in B_N) + 2\eta,$$  

(E.6)

where the equality follows from the law of total probability and the inequality follows from (EC.5). Let $\rho := \frac{(a+1)}{2\alpha} > 0$. Then, for all sufficiently large $N \in \mathbb{N}$:

$$\mathbb{P}^N \left(R_N \mathbb{P}(\xi \notin S_N) > 3\eta\right) \leq \mathbb{P}^N \left(R_N \mathbb{P}(\xi \notin S_N, \xi \in B_N) > \eta\right)$$

(E.7)

$$\leq \eta^{-1} R_N \mathbb{E}^{\mathbb{P}^N} \left[ \mathbb{P} \left(\xi \notin S_N, \xi \in B_N \mid \hat{\xi}^1, \ldots, \hat{\xi}^N\right) \right]$$

(E.8)

$$= \eta^{-1} R_N \mathbb{E}^{\mathbb{P}^N} \left[ \mathbb{P}^N \left(\xi \notin S_N, \xi \in B_N \mid \xi\right) \right]$$

(E.9)

$$= \eta^{-1} R_N \mathbb{E}^{\mathbb{P}^N} \left[ \mathbb{P} \left(\|\xi - \hat{\xi}\| > \epsilon_N, \xi \in B_N \mid \xi\right) \right]$$

(E.10)

$$= \eta^{-1} R_N \mathbb{E}^{\mathbb{P}^N} \left[ \mathbb{P} \left(\|\xi - \hat{\xi}\| > \epsilon_N, \xi \in B_N \mid \xi\right)^N \right]$$

(E.11)

$$\leq \eta^{-1} R_N \left(1 - \alpha_N \epsilon_N^d\right)^N$$

(E.12)

$$\leq \eta^{-1} R_N \exp \left(-N \alpha_N \epsilon_N^d\right)$$

(E.13)

$$\leq \eta^{-1} R_N \exp \left(-\kappa^d N \frac{\alpha(a+1)}{2\alpha} \alpha_N\right)$$

(E.14)

$$= \eta^{-1} R_N \exp \left(-\kappa^d \eta^d \left(\log N\right)^{1+p}\right).$$  

(E.15)
Line (EC.7) follows from (EC.6), (EC.8) follows from Markov’s inequality, (EC.9) follows from the law of iterated expectation, and (EC.10) follows from the definition of $S_N$. Line (EC.11) follows because, conditional on $\xi$, the random variables $\|\xi - \hat{\xi}^1\|, \ldots, \|\xi - \hat{\xi}^N\|$ are independent. Line (EC.12) follows from the definition of $B_N$, and (EC.13) follows from the mean value theorem. Line (EC.14) holds since Assumption 2 implies that $\epsilon_N \geq \kappa N^{-\frac{1}{d} + 1}$. Line (EC.15) follows from the definitions of $\alpha_N$, $R_N$, and $\rho$. Since $\rho > 0$, it follows from (EC.15) and the definition of $R_N$ that

$$\sum_{N=1}^{\infty} \mathbb{P}(R_N \mathbb{P}(\xi \notin S_N) > 3\eta) < \infty, \quad \forall \eta > 0,$$

and thus the Borel-Cantelli lemma implies that $R_N \mathbb{P}(\xi \notin S_N) \to 0$ as $N \to \infty$, $\mathbb{P}^\infty$-almost surely. □

Appendix C: Lower bound on Example 1

In Section 4.3, we introduced a lower bound $J^\bar{\cdot}$ on the optimal cost $J^*$ of Problem (1). In Theorem 2, we showed under mild assumptions that $J^\bar{\cdot}$ also provides an asymptotic lower bound on the optimal cost $\hat{J}_N$ of Problem (2). In this appendix, we will demonstrate the practical value of this asymptotic lower bound by revisiting the stochastic inventory problem from Example 1 in Section 3.

We recall that this stochastic inventory problem is given by

$$J^* = \minimize_{x_2: \mathbb{R} \rightarrow \mathbb{R}, x_3: \mathbb{R}^2 \rightarrow \mathbb{R}} \mathbb{E}[x_2(\xi_1) + 2x_3(\xi_1, \xi_2)]$$

subject to

$$x_2(\xi_1) + x_3(\xi_1, \xi_2) \geq \xi_1 + \xi_2 \quad \text{a.s.}$$

$$x_2(\xi_1), x_3(\xi_1, \xi_2) \geq 0 \quad \text{a.s.}$$

where the random variables $\xi = (\xi_1, \xi_2) \in \mathbb{R}^2$ denote the preorder and regular demand of a new product.

We assume throughout this appendix that this stochastic process satisfies Assumption 1 and is contained in $\Xi := \mathbb{R}^2_+$. In Appendix A, we showed that this stochastic inventory problem satisfies Assumption 3. Following the notation from Section 4.3, the lower bound for this problem is given by $J = \lim_{\rho \downarrow 0} J^\rho$, where

$$J^\rho := \minimize_{x, S \subseteq \Xi} \mathbb{E}\left[ (x_2(\xi_1) + 2x_3(\xi_1, \xi_2)) I \{ \xi \in S \} \right]$$

subject to

$$x_2(\xi_1) + x_3(\xi_1, \xi_2) \geq \xi_1 + \xi_2 \quad \forall \xi \in \hat{S}$$

$$x_2(\xi_1), x_3(\xi_1, \xi_2) \geq 0 \quad \forall \xi \in \hat{S}$$

$$\mathbb{P}(\xi \in \hat{S}) \geq 1 - \rho. \quad \text{(EC.16)}$$

In the remainder of this appendix, we prove the following proposition for the above example which, in combination with Theorem 2, shows that adding robustness to the historical data overcomes the overfitting phenomenon discussed in Section 3.

**Proposition EC.1.** $J^* = J$.

**Proof.** Our proof is split into several steps. First, we show that additional constraints can be added to (EC.16) without affecting its optimal cost. Indeed, we observe that the cost of procuring inventory is $1 per unit immediately after the preorder demand $\xi_1$ is observed, which is cheaper than producing inventory at $2
per unit after the regular demand $\xi_2$ is observed. Since all the demand will eventually need to be satisfied, we can without loss of generality restrict the decision rules in (EC.16) to those which satisfy $x_2(\zeta_1) \geq \zeta_1$ and $x_3(\zeta_1, \zeta_2) = \max\{0, \zeta_1 + \zeta_2 - x_2(\zeta_1)\}$ for every realization $\zeta \in \tilde{S}$. Therefore,

$$J_\rho = \begin{cases} 
\text{minimize} & \mathbb{E}\left[(x_2(\zeta_1) + 2 \max\{0, \zeta_1 + \zeta_2 - x_2(\zeta_1)\}) I\{\xi \in \tilde{S}\}\right] \\
\text{subject to} & x_2(\zeta_1) \geq \zeta_1 \quad \forall \zeta \in \tilde{S} \\
& \mathbb{P}\left(\xi \in \tilde{S}\right) \geq 1 - \rho 
\end{cases}$$

subject to $x_2(\zeta_1) \geq 0 \quad \forall \zeta \in \Xi$

$$\mathbb{P}\left(\xi \in \tilde{S}\right) \geq 1 - \rho,$$

(1) where the second equality follows from substituting $x_2(\zeta_1) = x_2(\zeta_1) - \zeta_1$. Notice that the constraints and objective function of (EC.17) only apply for realizations in $\tilde{S}$. In particular, we may add the constraint that $x_2(\zeta_1) \geq 0$ for all $\zeta \notin \tilde{S}$ to the above optimization problem without affecting its objective function. Thus, it follows from the aforementioned reasoning that

$$J_\rho = \begin{cases} 
\text{minimize} & \mathbb{E}\left[(\bar{x}_2(\zeta_1) + \zeta_1 + 2 \max\{0, \zeta_2 - \bar{x}_2(\zeta_1)\}) I\{\xi \in \tilde{S}\}\right] \\
\text{subject to} & \bar{x}_2(\zeta_1) \geq 0 \quad \forall \zeta \in \Xi \\
& \mathbb{P}\left(\xi \in \tilde{S}\right) \geq 1 - \rho 
\end{cases}$$

Next, we will develop a lower bound on the optimization problem in (EC.18). Indeed, for any feasible solution to the optimization problem in (EC.18), we observe that the function $\bar{x}_2(\zeta_1) + \zeta_2 + 2 \max\{0, \zeta_2 - \bar{x}_2(\zeta_1)\}$ is nonnegative almost surely. Therefore, for any arbitrary $r \geq 0$, a lower bound on $J_\rho$ is given by

$$J_{\rho,r} := \begin{cases} 
\text{minimize} & \mathbb{E}\left[(\bar{x}_2(\zeta_1) + \zeta_1 + 2 \max\{0, \zeta_2 - \bar{x}_2(\zeta_1)\}) I\{\xi \in \tilde{S}, \|\xi\| \leq r\}\right] \\
\text{subject to} & \bar{x}_2(\zeta_1) \geq 0 \quad \forall \zeta \in \Xi \\
& \mathbb{P}\left(\xi \in \tilde{S}\right) \geq 1 - \rho 
\end{cases}$$

Furthermore, the objective function in the above problem only considers realizations which satisfy $\zeta_2 \leq r$. Thus, we can impose the constraint that $\bar{x}_2(\zeta_1) \leq r$ for all $\zeta \in \Xi$ into (EC.19) without changing its optimal cost:

$$J_{\rho,L} := \begin{cases} 
\text{minimize} & \mathbb{E}\left[(\bar{x}_2(\zeta_1) + \zeta_1 + 2 \max\{0, \zeta_2 - \bar{x}_2(\zeta_1)\}) I\{\xi \in \tilde{S}, \|\xi\| \leq r\}\right] \\
\text{subject to} & 0 \leq \bar{x}_2(\zeta_1) \leq r \quad \forall \zeta \in \Xi \\
& \mathbb{P}\left(\xi \in \tilde{S}\right) \geq 1 - \rho 
\end{cases}$$

We now use Assumption 1 to obtain a lower bound on (EC.20). Indeed, Assumption 1 says that there exists an $a > 1$ such that $b := \mathbb{E}[\exp(\|\xi\|^a)] < \infty$. Therefore, for any feasible solution to the optimization problem in (EC.20),

$$\mathbb{E}\left[(\bar{x}_2(\zeta_1) + \zeta_1 + 2 \max\{0, \zeta_2 - \bar{x}_2(\zeta_1)\}) I\{\xi \notin \tilde{S} \text{ or } \|\xi\| > r\}\right]$$
The first inequality follows from the definition of $J$ from the definition of iterated expectation. The second inequality follows from removing constraints, and the final equality follows used in the proof of Theorem EC.2.

We now combine the above results to prove the main result. Indeed,

$$J = \lim_{\rho \to 0} J_\rho \geq \lim_{r \to \infty} \lim_{\rho \to 0} J_{\rho, r} \geq \lim_{\rho \to 0} \lim_{r \to \infty} -h(\rho, r) + J^* = J^*. $$

The first inequality follows because $J_\rho \geq J_{\rho, r}$ for any arbitrary $r \geq 0$ and the quantity $\lim_{r \to \infty} J_{\rho, r}$ is monotonically increasing in $r$. The second inequality follows from (EC.25). The final equality follows from the definition of $h(\rho, L)$. Since the inequality $J \leq J^*$ always holds, our proof is complete. □

Appendix D: Proof of Theorem 2 from Section 4.3

In this appendix, we present the proof of Theorem 2. In Appendix D.1, we develop a general bound between sample robust optimization and distributionally robust optimization with the 1-Wasserstein ambiguity set (Theorem EC.1). In Appendix D.2, we use that bound to establish a uniform convergence guarantee for sample robust optimization (Theorem EC.2). In Appendix D.3, we combine Theorem EC.2 and Theorem 1 to prove Theorem 2. In Appendix D.4, we include some intermediary and rather technical results which were used in the proof of Theorem EC.2.
D.1. A bound for sample and distributionally robust optimization

We begin with a general relationship between sample robust optimization and distributionally robust optimization using the 1-Wasserstein ambiguity set. Specifically, we show that sample robust optimization is, deterministically, nearly an upper bound on distributionally robust optimization with the 1-Wasserstein ambiguity set. Let \( \hat{N} \) denote the empirical distribution of historical data \( \hat{\xi}^1, \ldots, \hat{\xi}^N \in \mathbb{R}^d \), and \( Z \subseteq \mathbb{R}^d \) be any set that contains the historical data. We begin with an intermediary result which establishes a bound for the case where there is a single data point.

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

\[
\sup_{Q \in \mathcal{P}(Z) : \mathbb{E}_Q[\|\xi - \hat{\xi}\|] \leq \theta_1} \mathbb{E}_Q[f(\xi)] \leq \sup_{\zeta \in Z : \|\zeta - \hat{\xi}\| \leq \theta_2} f(\zeta) + \frac{2 \theta_1}{\theta_2} \sup_{\zeta \in Z} |f(\zeta)|. \tag{EC.26}
\]

**Proof.** First, we apply the Richter-Rogonsinski Theorem (see Theorem 7.32 and Proposition 6.40 of Shapiro et al. (2009)), which says that a distributionally robust optimization problem with \( m \) moment constraints is equivalent to optimizing a weighted average of \( m + 1 \) points. Thus,

\[
\sup_{Q \in \mathcal{P}(Z) : \mathbb{E}_Q[\|\xi - \hat{\xi}\|] \leq \theta_1} \mathbb{E}_Q[f(\xi)] = \sup_{0 \leq \lambda \leq 1} \left\{ \lambda \left( \sup_{\zeta \in Z : \|\zeta - \hat{\xi}\| \leq \frac{\theta_1}{\lambda}} f(\zeta) \right) + (1 - \lambda) \left( \sup_{\zeta \in Z : \|\zeta - \hat{\xi}\| \leq \frac{\theta_2}{1 - \lambda}} f(\zeta) \right) \right\}. \tag{EC.27}
\]

Let us assume from this point onward that \( \sup_{\zeta \in Z} |f(\zeta)| < \infty \); indeed, if \( \sup_{\zeta \in Z} |f(\zeta)| = \infty \), then the inequality in (EC.26) would trivially hold since the right-hand side would equal infinity. Then, it follows from (EC.27) that

\[
\sup_{Q \in \mathcal{P}(Z) : \mathbb{E}_Q[\|\xi - \hat{\xi}\|] \leq \theta_1} \mathbb{E}_Q[f(\xi)] = \sup_{0 \leq \lambda \leq 1} \left\{ \lambda \left( \sup_{\zeta \in Z : \|\zeta - \hat{\xi}\| \leq \frac{\theta_1}{\lambda}} f(\zeta) \right) + (1 - \lambda) \left( \sup_{\zeta \in Z : \|\zeta - \hat{\xi}\| \leq \frac{\theta_2}{1 - \lambda}} f(\zeta) \right) \right\}. \tag{EC.28}
\]

We observe that the optimization problem over \( \lambda \) is outer-most optimization problem on the right side of the equality in (EC.28) 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 that the right side of (EC.28) is equal to

\[
\sup_{0 \leq \lambda \leq 1 - \frac{\theta_1}{\theta_2}} \left\{ \lambda \left( \sup_{\zeta \in Z : \|\zeta - \hat{\xi}\| \leq \frac{\theta_1}{\lambda}} f(\zeta) \right) + (1 - \lambda) \left( \sup_{\zeta \in Z : \|\zeta - \hat{\xi}\| \leq \frac{\theta_2}{1 - \lambda}} f(\zeta) \right) \right\}. \tag{EC.29}
\]

Next, we observe that \( \frac{\theta_1}{\theta_2} \leq \theta_2 \) for every feasible \( \lambda \) for the above optimization problem. Using this inequality, we obtain the following upper bound:

\[
\sup_{Q \in \mathcal{P}(Z) : \mathbb{E}_Q[\|\xi - \hat{\xi}\|] \leq \theta_1} \mathbb{E}_Q[f(\xi)] \leq \sup_{0 \leq \lambda \leq 1 - \frac{\theta_1}{\theta_2}} \left\{ \lambda \left( \sup_{\zeta \in Z : \|\zeta - \hat{\xi}\| \leq \frac{\theta_1}{\lambda}} f(\zeta) \right) + (1 - \lambda) \left( \sup_{\zeta \in Z : \|\zeta - \hat{\xi}\| \leq \theta_2} f(\zeta) \right) \right\}
\]
Therefore, \( f \) is obtained by applying the global lower and upper bounds on Line (EC.31) follows because we can without loss of generality restrict \( \lambda \) to the interval \([0, \frac{\theta_1}{\theta_2}]\). Line (EC.32) is obtained by applying the global lower and upper bounds on \( f(\zeta) \). Finally, we obtain (EC.33) since

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

Combining (EC.30) and (EC.33), we obtain the desired result. \( \Box \)

We now present the main result of Appendix D.1, which extends the previous lemma to the general case with more than one data point.

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

\[
\sup_{Q \in \mathcal{P}(\mathcal{Z}) : d_1(Q, \bar{P}_N) \leq \theta_1} \mathbb{E}_Q[f(\xi)] \leq \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_{Q_j}[f(\xi)] + \frac{4\theta_1}{\theta_2} \sup_{\zeta \in \mathcal{Z}} |f(\zeta)|.
\]

**Proof.** We recall from the proof of Esfahani and Kuhn (2018, Theorem 4.2) that

\[
\left\{ \left. Q \in \mathcal{P}(\mathcal{Z}) \right| d_1(Q, \bar{P}_N) \leq \theta_1 \right\} = \left\{ \left. \frac{1}{N} \sum_{j=1}^{N} Q_j : \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_{Q_j}[\|\xi - \hat{\xi}_j\|] \leq \theta_1 \right\} \cap \{Q_1, \ldots, Q_N \in \mathcal{P}(\mathcal{Z})\}.
\]

Therefore,

\[
\sup_{Q \in \mathcal{P}(\mathcal{Z}) : d_1(Q, \bar{P}_N) \leq \theta_1} \mathbb{E}_Q[f(\xi)] = \sup_{\gamma \in \mathbb{R}_+^N} \left\{ \left. \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_{Q_j}[f(\xi)] : \frac{1}{N} \sum_{j=1}^{N} \gamma_j \leq \theta_1 \right\} \sup_{Q \in \mathcal{P}(\mathcal{Z}) : d_1(Q, \bar{P}_N) \leq \theta_1} \mathbb{E}_Q[f(\xi)].
\]
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,

$$\sup_{Q \in \mathcal{P}(z): d_1(Q, \tilde{P}) \leq \theta_1} \mathbb{E}_Q [f(\xi)] \leq \sup_{\gamma \in \mathbb{R}^N_+} \left\{ \frac{1}{N} \sum_{j \in [N]: 2\gamma_j \leq \theta_2} \sup_{Q \in \mathcal{P}(z): \|\xi - \xi^j\| \leq \gamma_j} \mathbb{E}_Q [f(\xi)] + \frac{1}{N} \sum_{j \in [N]: 2\gamma_j > \theta_2} \sup_{\xi \in \mathcal{Z}} |f(\xi)| : \frac{1}{N} \sum_{j=1}^N \gamma_j \leq \theta_1 \right\},$$

where the inequality follows from upper bounding each of the inner distributionally robust optimization problems for which $2\gamma_j > \theta_2$ by $\sup_{\xi \in \mathcal{Z}} |f(\xi)|$. 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 (EC.35) that

$$\sup_{Q \in \mathcal{P}(z): d_1(Q, \tilde{P}) \leq \theta_1} \mathbb{E}_Q [f(\xi)] \leq \sup_{\gamma \in \mathbb{R}^N_+} \left\{ \frac{1}{N} \sum_{j \in [N]: 2\gamma_j \leq \theta_2} \sup_{Q \in \mathcal{P}(z): \|\xi - \xi^j\| \leq \gamma_j} \mathbb{E}_Q [f(\xi)] : \frac{1}{N} \sum_{j=1}^N \gamma_j \leq \theta_1 \right\} + \frac{2\theta_1}{\theta_2} \sup_{\xi \in \mathcal{Z}} |f(\xi)|.$$ 

To conclude the proof, we apply Lemma EC.1 to each distributionally robust optimization problem in (EC.36) to obtain the following upper bounds.

$$\sup_{Q \in \mathcal{P}(z): d_1(Q, \tilde{P}) \leq \theta_1} \mathbb{E}_Q [f(\xi)] \leq \sup_{\gamma \in \mathbb{R}^N_+} \left\{ \frac{1}{N} \sum_{j \in [N]: 2\gamma_j \leq \theta_2} \left( \sup_{\xi \in \mathcal{Z}: \|\xi - \xi^j\| \leq \theta_2} f(\xi) + \frac{2\gamma_j}{\theta_2} \sup_{\xi \in \mathcal{Z}} |f(\xi)| \right) : \frac{1}{N} \sum_{j=1}^N \gamma_j \leq \theta_1 \right\} + \frac{2\theta_1}{\theta_2} \sup_{\xi \in \mathcal{Z}} |f(\xi)|$$

$$\leq \sup_{\gamma \in \mathbb{R}^N_+} \left\{ \frac{1}{N} \sum_{j=1}^N \left( \sup_{\xi \in \mathcal{Z}: \|\xi - \xi^j\| \leq \theta_2} f(\xi) + \frac{2\gamma_j}{\theta_2} \sup_{\xi \in \mathcal{Z}} |f(\xi)| \right) : \frac{1}{N} \sum_{j=1}^N \gamma_j \leq \theta_1 \right\} + \frac{2\theta_1}{\theta_2} \sup_{\xi \in \mathcal{Z}} |f(\xi)|$$

$$= \frac{1}{N} \sum_{j=1}^N \sup_{\xi \in \mathcal{Z}: \|\xi - \xi^j\| \leq \theta_2} f(\xi) + \frac{4\theta_1}{\theta_2} \sup_{\xi \in \mathcal{Z}} |f(\xi)|.$$ 

Line (EC.37) follows from applying Lemma EC.1 to (EC.36). Line (EC.38) follows because

$$\sup_{\xi \in \mathcal{Z}: \|\xi - \xi^j\| \leq \theta_2} f(\xi) + \frac{2\gamma_j}{\theta_2} \sup_{\xi \in \mathcal{Z}} |f(\xi)| \geq 0$$

for each component that satisfies $2\gamma_j > \theta_2$, and thus adding these quantities to (EC.37) results in an upper bound. Finally, (EC.39) follows from the constraint $\frac{1}{N} \sum_{j=1}^N \gamma_j \leq \theta_1$. This concludes the proof. \qed

**D.2. Uniform convergence of sample robust optimization**

In the previous Appendix D.1, we established a deterministic bound (Theorem EC.1) between sample robust optimization and distributionally robust optimization with the 1-Wasserstein ambiguity set. We will now combine that bound with a concentration inequality of Fournier and Guillin (2015) to establish a uniform convergence result for sample robust optimization (Theorem EC.2). Note that the proof in this section will employ Theorem 1 from Section 4.2 and Theorem EC.1 from Appendix D.1. For clarity of exposition, some intermediary and rather technical details of our proof are relegated to Appendix D.4.
Theorem EC.2. If Assumptions 1 and 2 hold, then there exists a \( \tilde{N} \in \mathbb{N} \), \( \mathbb{P}^\infty \)-almost surely, such that

\[
\mathbb{E} \left[ f(\xi) \mathbb{I} \{ \xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j \} \right] \leq \frac{1}{N} \sum_{j=1}^{\tilde{N}} \sup_{\xi \in U_N^j} f(\xi) + M_N \sup_{\xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j} |f(\xi)|,
\]

for all \( N \geq \tilde{N} \) and all measurable functions \( f : \mathbb{R}^d \to \mathbb{R} \), where \( M_N := N^{-\frac{1}{2}} \log N \to 0 \) as \( N \to \infty \).

Proof. Let \( \kappa > 0 \) be the coefficient from Assumption 2, and define \( \bar{\kappa} = \kappa / 8 \). For each \( N \in \mathbb{N} \), define

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

Then, it follows from Fournier and Guillin (2015) and Assumption 1 that \( d_1(\mathbb{P}, \hat{\mathbb{P}}_N) \leq \delta_N \) for all sufficiently large \( N \in \mathbb{N} \), \( \mathbb{P}^\infty \)-almost surely (see Lemma EC.2 in Appendix D.4). Therefore, for every measurable function \( f : \mathbb{R}^d \to \mathbb{R} \),

\[
\mathbb{E} \left[ f(\xi) \mathbb{I} \{ \xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j \} \right] = \mathbb{E} \left[ \left( f(\xi) + \sup_{\xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j} |f(\xi)| \right) \mathbb{I} \{ \xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j \} \right] - \mathbb{E} \left[ \left( \sup_{\xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j} |f(\xi)| \right) \mathbb{I} \{ \xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j \} \right] \mathbb{P}(\xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j),
\]

where the inequality holds for all sufficiently large \( N \in \mathbb{N} \), \( \mathbb{P}^\infty \)-almost surely. Next, we observe that \( g(\xi) \) equals zero when \( \xi \) is not an element of \( \bigcup_{j=1}^{\tilde{N}} U_N^j \) 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 \( \bigcup_{j=1}^{\tilde{N}} U_N^j \) (see Lemma EC.3 in Appendix D.4). Therefore, (EC.40) is equal to

\[
\mathbb{E} \left[ f(\xi) \mathbb{I} \{ \xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j \} \right] = \mathbb{E} \left[ \left( f(\xi) + \sup_{\xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j} |f(\xi)| \right) \mathbb{I} \{ \xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j \} \right] - \mathbb{E} \left[ \left( \sup_{\xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j} |f(\xi)| \right) \mathbb{I} \{ \xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j \} \right] \mathbb{P}(\xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j),
\]

where the first equality follows because the support of probability distributions in the outer-most supremum is restricted to those which assign measure only on \( \bigcup_{j=1}^{\tilde{N}} U_N^j \), and the second equality follows because \( \sup_{\xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j} |f(\xi)| \) 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 \( N \in \mathbb{N} \). Thus, it follows from Theorem EC.1 that (EC.41) is upper bounded by

\[
\frac{1}{N} \sum_{j=1}^{\tilde{N}} \sup_{\xi \in U_N^j} f(\xi) + \frac{4\delta_N}{\epsilon_N} \sup_{\xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j} |f(\xi)| + \left( \sup_{\xi \in \bigcup_{j=1}^{\tilde{N}} U_N^j} |f(\xi)| \right) \mathbb{P}(\xi \notin \bigcup_{j=1}^{\tilde{N}} U_N^j).
\]
By the definition of $\delta_N$, and since $\epsilon_N = \kappa N^{-\frac{d}{2}}$ when $d = 1$ and $\epsilon_N = \kappa N^{-\frac{1}{2}}$ when $d \geq 2$, we have that \( \frac{4 \delta_N}{\epsilon_N} \leq \frac{M}{2} \) for all sufficiently large $N$. Finally, Theorem 1 implies that $\mathbb{P}(\xi \notin \bigcup_{j=1}^{N} U_j) \leq \frac{M}{2}$ for all sufficiently large $N$, $\mathbb{P}^\infty$-almost surely. Combining (EC.40), (EC.41), and (EC.42), we obtain the desired result. \(\Box\)

**D.3. Proof of Theorem 2**

We now present the proof of Theorem 2. We first prove the lower bound in Theorem 2 by combining Theorem 1 from Section 4.2 and Theorem EC.2 from Appendix D.2.

**Theorem 2A.** Suppose Assumptions 1, 2, 3 hold. Then, $\mathbb{P}^\infty$-almost surely we have

\[
\hat{J} N \leq \liminf_{N \to \infty} \hat{J} N. \tag{EC.43}
\]

**Proof.** Recall from Assumption 1 that $b := \mathbb{E}[\exp(||\xi||^a)] < \infty$ for some $a > 1$, and let $L \geq 0$ be the constant from Assumption 3. Then,

\[
\sum_{N=1}^{\infty} \mathbb{P}^N \left( \sup_{\xi \in \bigcup_{j=1}^{N} U_j} L(1 + ||\xi||) > \log N \right) = \sum_{N=1}^{\infty} \mathbb{P}^N \left( \max_{j \in [N]} \left\{ L(1 + ||\xi^j|| + \epsilon_N) \right\} > \log N \right) \tag{EC.44}
\]

\[
\leq \sum_{N=1}^{\infty} N \mathbb{P} \left( L(1 + ||\xi|| + \epsilon_N) > \log N \right) \tag{EC.45}
\]

\[
= \sum_{N=1}^{\infty} N \mathbb{P} \left( ||\xi|| > \frac{\log N}{L} - 1 - \epsilon_N \right) \tag{EC.46}
\]

\[
= \sum_{N=1}^{\infty} N \mathbb{P} \left( \exp(||\xi||^a) > \exp \left( \left( \frac{\log N}{L} - 1 - \epsilon_N \right)^a \right) \right) \tag{EC.47}
\]

where (EC.44) follows from the definition of the uncertainty sets, (EC.45) follows from the union bound, (EC.46) follows from Markov’s inequality, and (EC.47) follows from $a > 1$. Therefore, the Borel-Cantelli lemma and Assumption 3 imply that the following equality holds for all sufficiently large $N \in \mathbb{N}$, $\mathbb{P}^\infty$-almost surely:

\[
\hat{J} N = \min_{x \in X} \frac{1}{N} \sum_{j=1}^{N} \max_{\xi \in \bigcup_{j=1}^{N} U_j} \sum_{t=1}^{T} c_t(\xi) \cdot x_t(\xi_1, \ldots, \xi_{t-1}) \tag{EC.48}
\]

subject to \( T \sum_{t=1}^{T} A_t(\xi) x_t(\xi_1, \ldots, \xi_{t-1}) \leq b(\xi) \quad \forall \xi \in \bigcup_{j=1}^{N} U_j \quad \forall \xi \in \bigcup_{j=1}^{N} U_j, t. \)

Moreover, since $c_1(\xi) \in \mathbb{R}^{n_1}, \ldots, c_T(\xi) \in \mathbb{R}^{n_T}$ are linear functions of the stochastic process, it follows from identical reasoning as (EC.44)-(EC.47) that $\sup_{\xi \in \bigcup_{j=1}^{N} U_j} ||c_t(\xi)|| \leq \log N$ for all sufficiently large $N \in \mathbb{N}$, $\mathbb{P}^\infty$-almost surely.
We now apply Theorem EC.2 to obtain an (asymptotic) lower bound on the optimization problem in (EC.48). Indeed, let \( M_N \) be shorthand for \( N^{-\frac{1}{(a+1)(b+1)}} \log N \). Then, for all sufficiently large \( N \in \mathbb{N} \), \( \mathbb{P}^\infty \)-almost surely, and for any decision rule which is feasible for the optimization problem in (EC.48),

\[
\mathbb{E} \left[ \sum_{t=1}^{T} c_t(\xi) \cdot x_t(\xi_1, \ldots, \xi_{t-1}) \left\| \xi \in \bigcup_{j=1}^{N} \mathcal{U}_N^j \right\| \right] \\
\leq \frac{1}{N} \sum_{j=1}^{N} \sup_{\zeta \in \mathcal{U}_N^j} \sum_{t=1}^{T} c_t(\zeta) \cdot x_t(\zeta_1, \ldots, \zeta_{t-1}) + M_N \sup_{\zeta \in \bigcup_{j=1}^{N} \mathcal{U}_N^j} \sum_{t=1}^{T} c_t(\zeta) \cdot x_t(\zeta_1, \ldots, \zeta_{t-1}) \\
\leq \frac{1}{N} \sum_{j=1}^{N} \sup_{\zeta \in \mathcal{U}_N^j} \sum_{t=1}^{T} c_t(\zeta) \cdot x_t(\zeta_1, \ldots, \zeta_{t-1}) + M_N \sum_{j=1}^{N} \sup_{\zeta \in \bigcup_{j=1}^{N} \mathcal{U}_N^j} \| c_t(\zeta) \| \| x_t(\zeta_1, \ldots, \zeta_{t-1}) \|, \\
\leq \frac{1}{N} \sum_{j=1}^{N} \sup_{\zeta \in \mathcal{U}_N^j} \sum_{t=1}^{T} c_t(\zeta) \cdot x_t(\zeta_1, \ldots, \zeta_{t-1}) + TM_N (\log N)^2.
\]

where the first inequality follows from Theorem EC.2, the second inequality follows from the triangle inequality and Cauchy-Schwarz inequality, and the third inequality follows because \( \| c_t(\zeta) \| \leq \log N \) and \( \| x_t(\zeta_1, \ldots, \zeta_{t-1}) \| \leq \log N \) for all sufficiently large \( N \in \mathbb{N} \) and realizations in the uncertainty sets. We remark that the above bound holds uniformly for all decision rules which are feasible for the optimization problem in (EC.48). Therefore, we have shown that the following inequality holds for all sufficiently large \( N \in \mathbb{N} \), \( \mathbb{P}^\infty \)-almost surely:

\[
\hat{J}_N + TM_N (\log N)^2 \geq \min_{\xi \in \mathcal{X}} \mathbb{E} \left[ \sum_{t=1}^{T} c_t(\xi) \cdot x_t(\xi_1, \ldots, \xi_{t-1}) \right] \\
\quad \left\{ \xi \in \bigcup_{j=1}^{N} \mathcal{U}_N^j \right\} \\
\text{subject to} \quad \sum_{t=1}^{T} A_t(\xi) x_t(\xi_1, \ldots, \xi_{t-1}) \leq b(\xi) \quad \forall \zeta \in \bigcup_{j=1}^{N} \mathcal{U}_N^j, \\
\| x_t(\zeta_1, \ldots, \zeta_{t-1}) \| \leq \log N \quad \forall \zeta \in \bigcup_{j=1}^{N} \mathcal{U}_N^j, t.
\]

Note that \( TM_N (\log N)^2 \) converges to zero as the number of sample paths \( N \) tends to infinity. Next, we obtain a lower bound on the right side of the above inequality by removing the last row of constraints and relaxing \( \bigcup_{j=1}^{N} \mathcal{U}_N^j \) to any set which contains the stochastic process with sufficiently high probability:

\[
\min_{\xi \in \mathcal{X}, \mathcal{S} \subseteq \mathcal{S}} \mathbb{E} \left[ \sum_{t=1}^{T} c_t(\xi) \cdot x_t(\xi_1, \ldots, \xi_{t-1}) \right] \left\{ \xi \in \mathcal{S} \right\} \\
\text{subject to} \quad \sum_{t=1}^{T} A_t(\xi) x_t(\xi_1, \ldots, \xi_{t-1}) \leq b(\xi) \quad \forall \zeta \in \mathcal{S}, \\
\mathbb{P} \left( \xi \in \mathcal{S} \right) \geq \mathbb{P} \left( \mathcal{S} \in \bigcup_{j=1}^{N} \mathcal{U}_N^j \right).
\]

Finally, for any fixed \( \rho \in (0, 1) \), it follows from Theorem 1 that \( \mathbb{P} \left( \xi \in \bigcup_{j=1}^{N} \mathcal{U}_N^j \cap \mathcal{S} \right) \geq 1 - \rho \) for all sufficiently large \( N \in \mathbb{N} \), \( \mathbb{P}^\infty \)-almost surely. Furthermore, we observe that \( TM_N (\log N)^2 \) converges to zero as \( N \) tends to
infinity. Therefore, we have shown that the following inequality holds, \( \mathbb{P}^\infty \)-almost surely:

\[
\liminf_{N \to \infty} \hat{J}_N \geq \min_{x \in X} \mathbb{E} \left[ \sum_{t=1}^{T} c_t(\xi) \cdot x_t(\xi_1, \ldots, \xi_{t-1}) \mathbb{I} \left\{ \xi \in \tilde{S} \right\} \right]
\]

subject to

\[
\sum_{t=1}^{T} A_t(\xi)x_t(\xi_1, \ldots, \xi_{t-1}) \leq b(\xi) \quad \forall \xi \in \tilde{S}
\]

\[
\mathbb{P} \left( \xi \in \tilde{S} \right) \geq 1 - \rho.
\]

Since the inequality holds true for every \( \rho \in (0, 1) \), and since the optimal cost of the right optimization problem is monotonically decreasing in \( \rho \), we obtain the desired result. \( \square \)

We now prove the upper bound in Theorem 2.

**Theorem 2b.** 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.** Consider any \( \rho > 0 \) such that there is a decision rule \( \bar{x} \in X \) which satisfies

\[
\sum_{t=1}^{T} A_t(\xi)\bar{x}_t(\xi_1, \ldots, \xi_{t-1}) \leq b(\xi) \quad \forall \xi \in \Xi : \text{dist}(\xi, S) \leq \rho.
\]  

(EC.50)

Indeed, if no such \( \rho > 0 \) existed, then \( \bar{J} = \infty \) and the desired result follows immediately. We recall from Assumption 2 that \( \epsilon_N \to 0 \) as \( N \to \infty \). Therefore,

\[
\limsup_{N \to \infty} \hat{J}_N \leq \limsup_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} \sup_{\|\xi - \xi^j\| \leq \epsilon_N} \sum_{t=1}^{T} c_t(\xi) \cdot \bar{x}_t(\xi_1, \ldots, \xi_{t-1})
\]

\[
\leq \lim_{k \to \infty} \limsup_{N \to \infty} \frac{1}{N} \sum_{j=1}^{N} \sup_{\|\xi - \xi^j\| \leq \epsilon_k} \sum_{t=1}^{T} c_t(\xi) \cdot \bar{x}_t(\xi_1, \ldots, \xi_{t-1})
\]

\[
= \lim_{k \to \infty} \mathbb{E} \left[ \sup_{\|\xi - \xi^j\| \leq \epsilon_k} \sum_{t=1}^{T} c_t(\xi) \cdot \bar{x}_t(\xi_1, \ldots, \xi_{t-1}) \right] \quad \mathbb{P}^\infty \text{-almost surely}
\]

\[
= \mathbb{E} \left[ \sum_{t=1}^{T} c_t(\xi) \cdot \bar{x}_t(\xi_1, \ldots, \xi_{t-1}) \right].
\]  

(EC.51)

The first inequality holds because the decision rule is feasible but possibly suboptimal for Problem (2) for all \( N \geq \min \{ \hat{N} : \epsilon_k \leq \rho \} \). 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 (see “Notation” in Section 1), and the final equality follows the definition of the local upper semicontinuous envelope. Since the set of decision rules which satisfy (EC.50) expands as \( \rho \downarrow 0 \), we conclude that the following holds \( \mathbb{P}^\infty \)-almost surely:

\[
\limsup_{N \to \infty} \hat{J}_N \leq \lim_{\rho \downarrow 0} \min_{x \in X} \mathbb{E} \left[ \sum_{t=1}^{T} c_t(\xi) \cdot x_t(\xi_1, \ldots, \xi_{t-1}) \right]
\]

subject to

\[
\sum_{t=1}^{T} A_t(\xi)x_t(\xi_1, \ldots, \xi_{t-1}) \leq b(\xi) \quad \forall \xi \in \Xi : \text{dist}(\xi, S) \leq \rho.
\]

This concludes the proof. \( \square \)
D.4. Miscellaneous results

We conclude Appendix D with some intermediary and technical lemmas which were used in the proof of Theorem EC.2 in Appendix D.2. The first lemma (Lemma EC.2) follows as a corollary of Fournier and Guillin (2015, Theorem 2) and is included for completeness.

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

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

for any fixed \( \bar{k} > 0 \). Then, \( d_1(\mathbb{P}, \hat{\mathbb{P}}_N) \leq \delta_N \) for all sufficiently large \( N \in \mathbb{N} \), \( \mathbb{P}^\infty \)-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 \( b := \mathbb{E}[\exp(||\xi||^a)] < \infty \). Thus, it follows from Fournier and Guillin (2015, Theorem 2) that there exist constants \( c_1, c_2 > 0 \) (which depend only on \( a, b, \) and \( d \)) such that for all \( N \geq \bar{N} \),

\[
\mathbb{P}^N \left( \frac{d_1(\mathbb{P}, \hat{\mathbb{P}}_N)}{\delta_N} > 1 \right) \leq \begin{cases} c_1 \exp \left( -c_2 N \delta_N^2 \right), & \text{if } d = 1, \\ c_1 \exp \left( -\frac{c_2 N \delta_N^2}{\log (2 + \bar{k}^{-2} N^{\frac{1}{2}})} \right), & \text{if } d = 2, \\ c_1 \exp \left( -c_2 N \delta_N^2 \right), & \text{if } d \geq 3. \end{cases}
\]

(EC.52)

First, suppose \( d = 1 \) and \( N \geq \bar{N} \). Then, it follows from the definition of \( \delta_N = \bar{k} N^{-\frac{1}{2}} \log N \) and (EC.52) that

\[
\mathbb{P}^N \left( \frac{d_1(\mathbb{P}, \hat{\mathbb{P}}_N)}{\delta_N} > 1 \right) \leq c_1 \exp \left( -c_2 N \delta_N^2 \right) = c_1 \exp \left( -c_2 \bar{k}^2 (\log N)^2 \right).
\]

Second, suppose \( d = 2 \) and \( N \geq \bar{N} \). Then, it follows from the definition of \( \delta_N = \bar{k} N^{-\frac{1}{2}} (\log N)^2 \) and (EC.52) that there exists some constant \( \bar{c} > 0 \) (which depends only on \( \bar{k} \) and \( c_2 \)) such that

\[
\mathbb{P}^N \left( \frac{d_1(\mathbb{P}, \hat{\mathbb{P}}_N)}{\delta_N} > 1 \right) \leq c_1 \exp \left( -\frac{c_2 N \delta_N^2}{\log (2 + \bar{k}^{-2} N^{\frac{1}{2}})} \right) = c_1 \exp \left( -\frac{c_2 \bar{k}^2 (\log N)^4}{\log (2 + \bar{k}^{-2} N^{\frac{1}{2}})} \right) \leq c_1 \exp \left( -c_2 \bar{k}^2 (\log N)^4 \right) \leq c_1 \exp \left( -\bar{c} (\log N)^4 \right).
\]

Third, suppose \( d \geq 3 \) and \( N \geq \bar{N} \). Then, it follows from the definition of \( \delta_N = \bar{k} N^{-\frac{1}{2}} (\log N)^2 \) and (EC.52) that

\[
\mathbb{P}^N \left( \frac{d_1(\mathbb{P}, \hat{\mathbb{P}}_N)}{\delta_N} > 1 \right) \leq c_1 \exp \left( -c_2 N \delta_N^2 \right) = c_1 \exp \left( -c_2 (\log N)^{2d} \right).
\]

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

\[
\sum_{N=1}^{\infty} \mathbb{P}^N \left( \frac{d_1(\mathbb{P}, \hat{\mathbb{P}}_N)}{\delta_N} > 1 \right) < \infty,
\]

and thus the desired result follows from the Borel-Cantelli lemma. \( \square \)
The second lemma (Lemma EC.3) demonstrates that restrictions may be placed on the support of an ambiguity set in distributionally robust optimization without loss of generality when the objective function is nonnegative.

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

$$
\sup_{Q \in \mathcal{P}(\Xi); d_4(P, \hat{P}_N) \leq \theta} \mathbb{E}_Q \left[ g(\xi) \mathbb{I} \{ \xi \in \mathcal{Z} \} \right] = \sup_{Q \in \mathcal{P}(\Xi); d_4(P, \hat{P}_N) \leq \theta} \mathbb{E}_Q \left[ g(\xi) \right].
$$

**Proof.** For compactness, define $\bar{g}(\zeta) := g(\zeta) \mathbb{I} \{ \zeta \in \mathcal{Z} \}$ for all $\zeta \in \Xi$. It readily follows from $\mathcal{Z} \subseteq \Xi$ that

$$
\sup_{Q \in \mathcal{P}(\Xi); d_4(P, \hat{P}_N) \leq \theta} \mathbb{E}_Q \left[ \bar{g}(\xi) \right] \geq \sup_{Q \in \mathcal{P}(\Xi); d_4(P, \hat{P}_N) \leq \theta} \mathbb{E}_Q \left[ g(\xi) \right] = \sup_{Q \in \mathcal{P}(\Xi); d_4(P, \hat{P}_N) \leq \theta} \mathbb{E}_Q \left[ g(\xi) \right],
$$

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)),

$$
\sup_{Q \in \mathcal{P}(\Xi); d_4(P, \hat{P}_N) \leq \theta} \mathbb{E}_Q \left[ \bar{g}(\xi) \right] = \begin{cases} 
\sup_{\zeta^1, \zeta^2 \in \Xi, \lambda^j \in [0, 1]} \frac{1}{N} \sum_{j=1}^{N} \left( \lambda^j \bar{g}(\zeta^1) + (1 - \lambda^j) \bar{g}(\zeta^2) \right) \\
\text{subject to} \frac{1}{N} \sum_{j=1}^{N} \left( \lambda^j \| \zeta^1 - \tilde{\zeta}^i \| + (1 - \lambda^j) \| \zeta^2 - \tilde{\zeta}^i \| \right) \leq \theta.
\end{cases}
$$

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

$$
\sup_{\zeta^1, \zeta^2 \in \mathcal{Z}, \lambda^j \in [0, 1]} \frac{1}{N} \sum_{j=1}^{N} \left( \lambda^j \bar{g}(\zeta^1) + (1 - \lambda^j) \bar{g}(\zeta^2) \right) \\
\text{subject to} \frac{1}{N} \sum_{j=1}^{N} \left( \lambda^j \| \zeta^1 - \tilde{\zeta}^i \| + (1 - \lambda^j) \| \zeta^2 - \tilde{\zeta}^i \| \right) \leq \theta,
$$

where we replaced the domain of $\zeta^1$ and $\zeta^2$ by $\mathcal{Z}$. We have thus shown that

$$
\sup_{Q \in \mathcal{P}(\Xi); d_4(P, \hat{P}_N) \leq \theta} \mathbb{E}_Q \left[ \bar{g}(\xi) \right] \leq \frac{1}{N} \sum_{j=1}^{N} \left( \lambda^j \bar{g}(\zeta^1) + (1 - \lambda^j) \bar{g}(\zeta^2) \right) + \eta
$$

$$
\leq \frac{1}{N} \sum_{j=1}^{N} \left( \lambda^j \bar{g}(\hat{\zeta}^1) + (1 - \lambda^j) \bar{g}(\hat{\zeta}^2) \right) + \eta \leq \sup_{Q \in \mathcal{P}(\Xi); d_4(P, \hat{P}_N) \leq \theta} \mathbb{E}_Q \left[ \bar{g}(\xi) \right] + \eta.
$$

Since $\eta > 0$ was chosen arbitrarily, we have shown the other direction. This concludes the proof. □

**Appendix E: Details for Example 2 from Section 4.3**

In this appendix, we provide the omitted technical details of Example 2 from Section 4.3. For convenience, we repeat the example:
Consider the following stochastic optimization problem:

\[
\begin{align*}
\text{minimize} & \quad x_1 \\
\text{subject to} & \quad x_1 \geq \xi_1 \quad \text{a.s.,}
\end{align*}
\]

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 \(\Xi = [0, 2]\). We consider applying Problem (2) to the above optimization problem with a robustness parameter \(\epsilon_N = N^{-\frac{k}{4}}\). Suppose the true probability distribution is defined by \(P(\xi_1 > \alpha) = (1 - \alpha)^k\) for some fixed \(k > 0\) and every \(\alpha \in [0, 1]\). In this case, the true support of the underlying distribution is \(S = [0, 1]\) for any \(k > 0\). Then, employing the Hewitt-Savage zero-one law, we prove in Appendix E that the bounds in Theorem 2 are tight under different choices of \(k\):

<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 &lt; \limsup</em>{N \to \infty} \hat{J}_N &lt; \bar{J} \right) = 1 )</td>
</tr>
</tbody>
</table>

We now prove the above bounds. To begin, we recall that \(\mathbb{P}(\xi_1 > \alpha) = (1 - \alpha)^k\). Thus, for any \(k > 0\),

\[
\begin{align*}
J &= \lim_{\rho \downarrow 0} \min_{x_1 \in \mathbb{Z}} \{ x_1 : \mathbb{P}(x_1 \geq \xi_1) \geq 1 - \rho \} = 1, \quad \text{and} \\
\bar{J} &= \lim_{\rho \downarrow 0} \min_{x_1 \in \mathbb{Z}} \{ x_1 : x_1 \geq 1 + \rho \} = 2.
\end{align*}
\]

Furthermore, given historical data, the choice of the robustness parameter \(\epsilon_N = N^{-\frac{k}{4}}\), and \(\Xi = [0, 2]\),

\[
\hat{J}_N = \min_{x_1 \in \mathbb{Z}} \left\{ x_1 : x_1 \geq \xi_1, \forall \xi_i \in \bigcup_{j=1}^N U_j \right\} = \begin{cases} 1, & \text{if } \max_{j \in [N]} \xi_j \leq 1 - N^{-\frac{k}{4}}, \\ 2, & \text{if } \max_{j \in [N]} \xi_j > 1 - N^{-\frac{k}{4}}. \end{cases}
\]

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,

\[
\begin{align*}
&\mathbb{P}^\infty \left( \limsup_{N \to \infty} \hat{J}_N = 1 \right) \\
&= \mathbb{P}^\infty \left( \max_{j \in [N]} \xi_j^\hat{1} \leq 1 - N^{-\frac{k}{4}} \text{ for all sufficiently large } N \right) \\
&= \lim_{N \to \infty} \mathbb{P}^\infty \left( \max_{j \in [n]} \xi_j^\hat{1} \leq 1 - n^{-\frac{k}{4}} \text{ for all } n \geq N \right) \\
&= \lim_{N \to \infty} \mathbb{P}^\infty \left( \max_{j \in [N]} \xi_j^\hat{1} \leq 1 - N^{-\frac{k}{4}} \text{ and } \max_{j \in [n]} \xi_j^\hat{1} \leq 1 - n^{-\frac{k}{4}} \text{ for all } n \geq N + 1 \right)
\end{align*}
\]
Indeed, therefore, it follows from the Borel-Cantelli lemma that

\[
= \lim_{N \to \infty} P^N \left( \max_{j \in [N]} \hat{\xi}_j^3 \leq 1 - N^{-\frac{1}{2}} \right) \prod_{n=N+1}^{\infty} P \left( \max_{j \in [n]} \hat{\xi}_j^3 \leq 1 - n^{-\frac{1}{2}} \right) \quad \text{(EC.53)}
\]

\[
= \lim_{N \to \infty} P^N \left( \max_{j \in [N]} \hat{\xi}_j^3 \leq 1 - N^{-\frac{1}{2}} \right) \prod_{n=N+1}^{\infty} P \left( \max_{j \in [n]} \hat{\xi}_j^3 \leq 1 - (n-1)^{-\frac{1}{2}} \right) \quad \text{(EC.54)}
\]

\[
= \lim_{N \to \infty} P \left( \xi_1 \leq 1 - N^{-\frac{1}{2}} \right) \prod_{n=N+1}^{\infty} P \left( \xi_1 \leq 1 - n^{-\frac{1}{2}} \right) \quad \text{(EC.55)}
\]

\[
= \lim_{N \to \infty} \left( 1 - N^{-\frac{1}{2}} \right)^N \prod_{n=N+1}^{\infty} \left( 1 - n^{-\frac{1}{2}} \right). \quad \text{(EC.56)}
\]

Line (EC.53) follows from the law of total probability. Line (EC.54) follows because, conditional on \( \max_{j \in [n-1]} \hat{\xi}_j^3 \leq 1 - (n-1)^{-\frac{1}{2}} \), we have that \( \hat{\xi}_j^3 \leq 1 - n^{-\frac{1}{2}} \) for all \( j \in [n-1] \). Line (EC.55) follows from the independence of \( \hat{\xi}_j^3, j \in \mathbb{N} \). By evaluating the limit in (EC.56), we conclude the proof of Claim 1.

Next, we show that

\[
P^\infty \left( \liminf_{N \to \infty} \hat{J}_N = 1 \right) = 1 \text{ if } k \geq 3.
\]

Claim 2

Indeed,

\[
P^\infty \left( \liminf_{N \to \infty} \hat{J}_N = 1 \right) = P^\infty \left( \max_{j \in [N]} \hat{\xi}_j^3 \leq 1 - N^{-\frac{1}{2}} \text{ for infinitely many } N \right)
\]

\[
= \lim_{N \to \infty} P^\infty \left( \max_{j \in [N]} \hat{\xi}_j^3 \leq 1 - n^{-\frac{1}{2}} \text{ for some } n \geq N \right)
\]

\[
\geq \lim_{N \to \infty} P^N \left( \max_{j \in [N]} \hat{\xi}_j^3 \leq 1 - N^{-\frac{1}{2}} \right)
\]

\[
= \lim_{N \to \infty} P \left( \xi_1 \leq 1 - N^{-\frac{1}{2}} \right)^N \quad \text{(EC.57)}
\]

\[
= \lim_{N \to \infty} \left( 1 - N^{-\frac{1}{2}} \right)^N. \quad \text{(EC.58)}
\]

Line (EC.57) follows from the independence of \( \hat{\xi}_j^3, j \in \mathbb{N} \). We observe that the limit in (EC.58) 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 \( \{ \max_{j \in [N]} \hat{\xi}_j^3 \leq 1 - N^{-\frac{1}{2}} \text{ for infinitely many } N \} \) happens with probability zero or one. Thus, (EC.58) 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

\[
P^\infty \left( \liminf_{N \to \infty} \hat{J}_N = 1 \right) = 0 \text{ if } 0 < k < 3.
\]

Claim 3

Indeed, suppose that \( 0 < k < 3 \). Then,

\[
\sum_{N=1}^{\infty} P^\infty \left( \hat{J}_N = 1 \right) = \sum_{N=1}^{\infty} P^N \left( \max_{j \in [N]} \hat{\xi}_j^3 \leq 1 - N^{-\frac{1}{2}} \right) = \sum_{N=1}^{\infty} \left( 1 - N^{-\frac{1}{2}} \right)^N < \infty.
\]

Therefore, it follows from the Borel-Cantelli lemma that

\[
P^\infty \left( \liminf_{N \to \infty} \hat{J}_N = 1 \right) = P^\infty \left( \max_{j \in [N]} \hat{\xi}_j^3 > 1 - N^{-\frac{1}{2}} \text{ for all sufficiently large } N \right) = 0,
\]
when \(0 < k < 3\), which proves Claim 3.

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

**Appendix F: An Iterative Approach for Selecting Partitions**

In this appendix, we describe the iterative algorithm for selecting partitions for finite adaptability in Section 5.2. The algorithm is an extension and refinement of ideas of Bertsimas and Dunning (2016) and Postek and Hertog (2016), which were designed for robust optimization. The performance of the algorithm is demonstrated in Section 7.

In a nutshell, the iterative algorithm can be described as follows. We start each iteration with a partition \(\mathcal{P} = \{P_1, \ldots, P_K\}\) of the set \(\Xi\), and we solve the resulting approximation of Problem (2) with finite adaptability. Then, for each region \(k \in [K]\) and each uncertainty set \(j \in [N]\), we find the realizations \(\zeta\) contained in \(U_j \cap P_k\) that are active for the previous decision rule, and recursively split the previous partition into a refined partition \(\mathcal{P} = \{\bar{P}_1, \ldots, \bar{P}_K\}\) by separating the active realizations. As established by the previous literature, this heuristic for creating partitions has theoretical justification for the case of robust optimization, and in this appendix we will extend this justification to the present setting of Problem (2).

The purpose of this appendix is to adapt the partitioning algorithm and theoretical justification from Bertsimas and Dunning (2016) and Postek and Hertog (2016) to Problem (2). Indeed, since Problem (2) involves averaging over the worst-case costs from multiple uncertainty sets, the algorithms and theory from existing literature require modification to be tractable and justified in the present setting.

This appendix is organized as follows. In Appendix F.1, we generalize the finite adaptability approach from Section 5.2 to the setting in which the vectors \(c_t(\xi)\) and matrices \(A_t(\xi)\) depend on the stochastic process. In Appendix F.2, we present the notation and theoretical justification for the partitioning heuristic in the context of Problem (2). In Appendix F.3, we present the partitioning algorithm in greater detail and discuss its tractability.

### F.1. General model for finite adaptivity

We begin by discussing finite adaptability in the general case of Problem (2). Let \(\mathcal{P} = \{P^1, \ldots, P^K\}\) denote a partition of the set \(\Xi\). We consider approximating Problem (2) by restricting its decision rules to those of the form

\[
x_t(\zeta_1, \ldots, \zeta_{t-1}) = \begin{cases} 
    x^1_t, & \text{if } (\zeta_1, \ldots, \zeta_{t-1}) \in P^1_{t-1}, \\
    \vdots & \\
    x^K_t, & \text{if } (\zeta_1, \ldots, \zeta_{t-1}) \in P^K_{t-1},
\end{cases}
\]

where \(P^k_t \subseteq \mathbb{R}^{d_1 + \cdots + d_t}\) is the projection of the region \(P^k\) onto the first \(t\) stages components. In contrast to the setting described in Section 5.2, we consider using static decision rules in each region as opposed to linear decision rules. We will revisit this distinction shortly.
After restricting to decision rules of the above form, we obtain an upper bound approximation of Problem (2) by solving the following optimization problem:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{N} \sum_{j=1}^{N} \max_{k \in K_j} \sup_{\zeta \in \mathcal{U}_k \cap P^k} \sum_{t=1}^{T} c_t(\zeta) x_t^k \\
\text{subject to} & \quad \sum_{t=1}^{T} A_t(\zeta)x_t^k \leq b(\zeta) \quad \forall \zeta \in \bigcup_{j=1}^{N} \mathcal{U}_j \cap P^k, \ k \in [K] \\
& \quad x_t^k = x_t^{k'}, \quad \forall (k, k', t) \in \mathcal{T}(\mathcal{P}),
\end{align*}
\]  

(EC.59)

where the decision variables are \(x_t^k \in X_t\) for each \(t \in [T]\) and \(k \in [K]\), and we use the shorthand \(\mathcal{T}(\mathcal{P}) := \mathcal{T}(P^1, \ldots, P^K)\) to denote the set of tuples \((k, k', t)\) which satisfy the conditions described in Proposition 1 from Section 5.2 to ensure that the decision rules are not anticipative. In the following we use \(a_{i,t}(\zeta) \in \mathbb{R}^{n_t}\) to denote the \(i\)th row of matrix \(A_t(\zeta) \in \mathbb{R}^{m \times n_t}\), and \(b_t(\zeta) \in \mathbb{R}^m\) to denote the \(i\)th component of vector \(b(\zeta) \in \mathbb{R}^m\).

In contrast to Section 5.2, we focus on finite adaptability with static decision rules due to computational tractability. Indeed, when \(A_t(\zeta)\) and \(b(\zeta)\) can depend on the stochastic process, it is known that linear decision rules do not remain computationally tractable in general. Moreover, we will now show that Problem (EC.59) is, in fact, a generalization of Problem (5) from Section 5.2.

**Proposition EC.2.** Problem (EC.59) encompasses Problem (5) as a special case.

**Proof.** For notational convenience, let \(\text{vec}(M)\) denote the vectorization of a matrix \(M\) into a single column vector, let \([M_1, \ldots, M_t]\) denote the horizontal concatenation of matrices \(M_1, \ldots, M_t\) into a single matrix, and let \(\tilde{\zeta}_t \in \mathbb{R}^{\sum_{i=1}^{d_t+1}}\) be the vertical concatenation of the scalar 1 and the vectors \(\zeta_1, \ldots, \zeta_{t-1}\) into a single vector. Using this notation, and starting from any instance of Problem (5), we can construct an equivalent instance of Problem (EC.59) using the following transformation. Indeed, for each \(t \in [T]\), \(k \in [K]\), and \(\zeta \in \Xi\), we define the following terms:

\[
\begin{align*}
x_t^k & := \text{vec}\left([x_{t,0}^k, X_{t,1}^k, \ldots, X_{t,t-1}^k]\right) \in \mathbb{R}^{n_t(\sum_{i=1}^{d_t+1})}, \ k \in [K], \\
c_t(\zeta) & := \text{vec}\left(\tilde{\zeta}_t c_{t,\tilde{c}_t}^T\right), \\
a_{i,t}(\zeta) & := \text{vec}\left(\tilde{\zeta}_t a_{i,\tilde{a}_t}^T\right), \ i \in [m],
\end{align*}
\]

where \(a_{i,t}\) is the \(i\)th row of matrix \(A_t\) used in Problem (5). Therefore, we readily observe that for all \(t \in [T]\) and \(k \in [K]\), the following equivalences hold:

\[
\begin{align*}
c_t(\zeta) \cdot x_t^k & \equiv c_t \cdot \left(x_{t,0}^k + \sum_{s=1}^{t-1} X_{t,s}^k \zeta_s\right) \\
a_{i,t}(\zeta) x_t^k & \equiv a_{i,t} \cdot \left(x_{t,0}^k + \sum_{s=1}^{t-1} X_{t,s}^k \zeta_s\right), \ i \in [m],
\end{align*}
\]

and the equivalence between Problem (5) and Problem (EC.59) is obtained. □

Finally, under Assumption 5, we observe that Problem (EC.59) can be reformulated as a finite-dimensional linear optimization problem using identical reformulation techniques as Theorems 3 and 4. In the remainder
of Appendix F, we will focus on developing a partitioning algorithm for Problem (EC.59), which immediately applies to the setting of Section 5.2 by applying Proposition EC.2.

F.2. Refined partitions

We next discuss some theoretical concepts about partitions in finite adaptability, which provide justification for the algorithm in the following section. For compactness, we henceforth let $z(\mathcal{P})$ denote the optimal cost of Problem (EC.59) with partition $\mathcal{P}$. Thus, using epigraph constraints, Problem (EC.59) with partition $\mathcal{P}$ has the form

$$z(\mathcal{P}) = \min \frac{1}{N} \sum_{j=1}^{N} v_j$$

subject to

$$\sum_{t=1}^{T} c_t(\zeta) \cdot x^k_t \leq v_j \quad \forall \zeta \in U_j \cap P^j, \, j \in [N], \, k \in K_j$$

$$\sum_{t=1}^{T} A_t(\zeta)x^k_t \leq b_\zeta(\zeta) \quad \forall \zeta \in \bigcup_{j=1}^{N} U_j \cap P^j, \, k \in [K]$$

$$x^k_t = x^{k'}_t \quad \forall (k, k', t) \in T(\mathcal{P})$$

where the decision variables are $v \in \mathbb{R}^N$ and $x^k \equiv (x^k_1, \ldots, x^k_T)$ for each $k \in [K]$. Our discussion on partitions in finite adaptability begins with the following definition.

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

Intuitively speaking, a refined partition $\mathcal{P}'$ is constructed by splitting some of the regions $P^k \in \mathcal{P}$. For a visualization of refined partitions, see Figure EC.1. The following intuitive proposition shows that the optimal cost of Problem (EC.60) with a refined partition is never worse than the optimal cost of Problem (EC.60) with the original partition.

**Proposition EC.3.** 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.** Let $x^1, \ldots, x^K$ and $v \in \mathbb{R}^N$ be optimal for Problem (EC.60) with partition $\mathcal{P}$. Then, we can construct a feasible (but possibly suboptimal) solution $\bar{x}^1, \ldots, \bar{x}^{\bar{K}}$ and $\bar{v} \in \mathbb{R}^N$ for Problem (EC.60) with the refined partition $\mathcal{P}'$ by setting $\bar{v} = v$ and $\bar{x}^l = x^k$ for each $l \in [\bar{K}]$ and $k \in [K]$ such that $\bar{P}^l \subseteq P^k$. Since this solution is possibly suboptimal for Problem (EC.60) with the refined partition $\mathcal{P}'$, it follows that $z(\mathcal{P}') \leq z(\mathcal{P})$. □

The above result establishes that a refined partition will never result in a worse optimal cost. On the other hand, a refined partition may not necessarily result in a strictly improved optimal cost. Thus, we desire a characterization of when a refined partition can lead to an improvement in the optimal cost. Towards that end, we introduce some further notation. Let $x^1, \ldots, x^K$ and $v \in \mathbb{R}^N$ be an optimal solution of Problem (EC.60) with a partition $\mathcal{P} = \{P^1, \ldots, P^K\}$ of $\Xi$. For each region $P^k \in \mathcal{P}$, we define the corresponding active set of constraints (cuts) by
We now show that these constraints involving $x$ in the active sets can be removed without impacting the optimal cost. That is, the nonanticipativity constraints of Problem (EC.61), i.e.,

Next, consider any $k,k' \in [K]$, $l,l' \in [K]$ and $t \in [T]$ such that $\mathcal{A}(P^k),\mathcal{A}(P^l) \neq \emptyset$, $\mathcal{A}(P^k) \subseteq P^l$, $\mathcal{A}(P^l) \subseteq P^{l'}$, $(k,k',t) \in \mathcal{T}(\mathcal{P})$ and $(l,l',t) \notin \mathcal{T}(\mathcal{P})$

**Proposition EC.4.** Let $\mathcal{P} \equiv \{P^1,\ldots,P^K\}$ be a partition of $\Xi$, and let $\mathcal{P} \equiv \{\bar{P}^1,\ldots,\bar{P}^K\}$ be a refined partition of $\mathcal{P}$. If $z(\mathcal{P}) < z(\mathcal{P})$, then at least one of the following two conditions hold:

(a) There exists a $k \in [K]$ such that $\mathcal{A}(P^k) \notin \bar{P}_l$ for all $l \in [K]$.

(b) There exists $k,k' \in [K]$, $l,l' \in [K]$ and $t \in [T]$ such that $\mathcal{A}(P^k),\mathcal{A}(P^l) \neq \emptyset$, $\mathcal{A}(P^k) \subseteq P^l$, $\mathcal{A}(P^l) \subseteq P^{l'}$, $(k,k',t) \in \mathcal{T}(\mathcal{P})$ and $(l,l',t) \notin \mathcal{T}(\mathcal{P})$.

**Proof.** Let $\mathcal{P} \equiv \{P^1,\ldots,P^K\}$ be a partition of $\Xi \subseteq \mathbb{R}^d$ and $\mathcal{P} \equiv \{\bar{P}^1,\ldots,\bar{P}^K\}$ be a refined partition of $\mathcal{P}$. First, we observe that all of the constraints in Problem (EC.60) which do not correspond to realizations in the active sets can be removed without impacting the optimal cost. That is,

$$z(\mathcal{P}) = \min \frac{1}{N} \sum_{j=1}^{N} v_j$$

subject to

$$\sum_{t=1}^{T} c_t(\zeta) \cdot x^k_t \leq v_j \quad \forall \zeta \in U^k, \mathcal{A}(P^k), j \in [N], k \in [K]$$

$$\sum_{t=1}^{T} A_t(\zeta)x^k_t \leq b(\zeta) \quad \forall \zeta \in \mathcal{A}(P^k), k \in [K]$$

$$x^k_t = x^{k'}_t, \quad \forall (k,k',t) \in \mathcal{T}(\mathcal{P}).$$

Next, consider any $k \in [K]$ for which $\mathcal{A}(P^k) = \emptyset$. In this case, the decision variables $x^k$ appear only in the nonanticipativity constraints of Problem (EC.61), i.e., the constraints of the form

$$x^k_t = x^{k'}_t, \quad \forall (k,k',t) \in \mathcal{T}(\mathcal{P}).$$

We now show that these constraints involving $x^k$ can be removed without impacting the optimal objective value. First, we observe that the set $\mathcal{T}(\mathcal{P})$ is transitive, in the sense that $(k,l,t),(k,s,t) \in \mathcal{T}(\mathcal{P})$ implies

$$\mathcal{A}(P^k) \subseteq \mathcal{A}(P^l) \subseteq \mathcal{A}(P^s)$$

for all $(k,l,t),(k,s,t) \in \mathcal{T}(\mathcal{P})$.
(l, s, t) ∈ T(\mathcal{P}). Therefore, removing the constraints involving x^k from Problem (EC.61) does not result in eliminating nonanticipativity constraints on the other variables in the problem. Thus, letting J := \{k ∈ [K]: \mathcal{A}(P^k) \neq \emptyset\}, we have shown that

\[ z(\mathcal{P}) = \min \frac{1}{N} \sum_{j=1}^{N} v_j \]

subject to

\[ \sum_{t=1}^{T} c_t(\zeta) \cdot x^k_t \leq v_j, \quad \forall \zeta \in U^t_j \cap \mathcal{A}(P^k), j \in [N], k \in J \]

\[ \sum_{t=1}^{T} A_i(\zeta) x^k_t \leq b(\zeta), \quad \forall \zeta \in \mathcal{A}(P^k), k \in J \]

\[ x^k_t = x^{k'}_{t'}, \quad \forall (k, k', t) \in T(\mathcal{P}), k, k' \in J. \]

We now prove the main result. Let \( \tilde{\mathcal{P}} \equiv \{\tilde{P}^1, \ldots, \tilde{P}^K\} \) be a refined partition of \( \mathcal{P} \) which satisfies neither Conditions (a) nor (b) in Proposition EC.4. Since Condition (a) does not hold, there exists a mapping \( \iota: J \rightarrow [K] \) such that \( \mathcal{A}(P^k) \subseteq \tilde{P}^{(k)} \subseteq P^k \) for every \( k \in J \). Since Condition (b) does not hold, \( (\iota(k), \iota(k'), t) \in T(\tilde{\mathcal{P}}) \) whenever \( k, k' \in J \) and \( (k, k', t) \in T(\mathcal{P}) \). Thus, the following is a relaxation of Problem (EC.60) with the refined partition \( \tilde{\mathcal{P}} \):

\[ \tilde{z}(\tilde{\mathcal{P}}) := \min \frac{1}{N} \sum_{j=1}^{N} v_j \]

subject to

\[ \sum_{t=1}^{T} c_t(\zeta) \cdot x^{(k)}_t \leq v_j, \quad \forall \zeta \in U^t_j \cap \mathcal{A}(P^k), j \in [N], k \in J \]

\[ \sum_{t=1}^{T} A_i(\zeta) x^{(k)}_t \leq b(\zeta), \quad \forall \zeta \in \mathcal{A}(P^k), k \in J \]

\[ x^{(k)}_t = x^{(k')}_{t'}, \quad \forall (k, k', t) \in T(\mathcal{P}), k, k' \in J. \]

However, we observe that Problem (EC.64) is identical to Problem (EC.63). Thus, we have shown that

\[ z(\mathcal{P}) = \tilde{z}(\tilde{\mathcal{P}}) \leq z(\tilde{\mathcal{P}}). \]

We conclude from Proposition EC.3 that \( z(\mathcal{P}) = z(\tilde{\mathcal{P}}) \) if Conditions (a) and (b) do not hold. Since we have shown the contrapositive of the desired result, the proof is complete. □

In Appendix F.3, we will use Proposition EC.4 to design the iterative algorithm for finding high-quality partitions for Problem (EC.60). In the remainder of the present Appendix F.2, we provide a further discussion on the conditions in Proposition EC.4. First, we remark that a similar condition to Condition (a) of Proposition EC.63 is found in Bertsimas and Dunning (2016, Theorem 1). Second, in order to show that Condition (b) of Proposition EC.63 is an alternative necessary condition in our setting for improving the optimal cost, we provide the following example.

**Example EC.1.** Consider a problem with \( T = 3, X = \mathbb{R}^3 \), and the problem parameters are given by

\[ A_1(\zeta) = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, A_2(\zeta) = \begin{bmatrix} 2 + \zeta_1 \\ 0 \\ 0 \end{bmatrix}, A_3(\zeta) = \begin{bmatrix} -\zeta_2 \\ -1 \\ -1 \end{bmatrix}, b(\zeta) = \begin{bmatrix} 0 \\ \frac{-\zeta_1 + \zeta_2}{2} \\ \frac{-\zeta_1 + \zeta_2}{2} \end{bmatrix}, c_1(\zeta) = 0, c_2(\zeta) = -1 - \zeta_1, c_3(\zeta) = 2\zeta_2. \]
Suppose there is no uncertainty after the third-stage (i.e., \( \xi \equiv (\xi_1, \xi_2) \in \mathbb{R}^2 \)), \( N = 1 \), and \( \Xi = \mathcal{U}_1 = [-1,1]^2 \). Finally, suppose we start with a partition \( \mathcal{P} = \{P^1, P^2\} \) of the form

\[
P^1 = \left\{ \zeta \in \Xi : \zeta_2 \geq \frac{1}{2} \right\}; \quad P^2 = \left\{ \zeta \in \Xi : \zeta_2 \leq \frac{1}{2} \right\}.
\]

Hence, the associated non-anticipatory constraints are \( \mathcal{T}(\mathcal{P}) = \{(1,2,1), (1,2,2)\} \). The resulting instance of Problem (EC.60) is given by

\[
z(\mathcal{P}) = \min_{x^1, x^2, v \in \mathbb{R}} v
\]

subject to

\[
\begin{align*}
(1-x_2^2) & \leq v \quad \forall \zeta \in P^k, k \in [2] \\
(2-x_1) & \leq v \quad \forall \zeta \in P^k, k \in [2] \\
\zeta_1 + \frac{1}{2} & \leq x_3 \\
\zeta_2 + \frac{1}{2} & \leq x_3 \\
\frac{1}{2} & \leq x_3
\end{align*}
\]

Since \( P^k = P^k \cap \mathcal{U}_1 \) are bounded hyperrectangles, the active realizations of each constraint are found at the extreme points of \( P^1 \) and \( P^2 \). Thus, plugging in the extreme points of \( P^k \) for each constraint involving \( x^k \), we reformulate the above optimization problem as

\[
z(\mathcal{P}) = \min_{x_2, x_3, v \in \mathbb{R}} v
\]

subject to

\[
\begin{align*}
\max\{-2x_2, 0\} & + \max\{x_1, 2x_1, x_2, -2x_2\} & \leq v \\
\max\{3x_2, x_2\} & + \max\{-x_3, -x_3, -x_3, -x_3\} & \leq 0 \\
x_2^1, x_3^1 & \geq 1.
\end{align*}
\]

We observe that the optimal cost of the above optimization problem is 4, and the optimal decisions are \( x^1_2 = -1 \) and \( x^1_3 = x^2_3 = 1 \). 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.4.

We now construct a refined partition that satisfies Condition (b) which improves the optimal objective value. Let the refined partition be defined as \( \mathcal{P} = \{\bar{P}^1, \bar{P}^2, \bar{P}^3, \bar{P}^4\} \), where

\[
\bar{P}^1 = \left\{ \zeta \in \Xi : \zeta_1 \geq 0, \zeta_2 \geq \frac{1}{2} \right\} ; \quad \bar{P}^2 = \left\{ \zeta \in \Xi : \zeta_1 \leq 0, \zeta_2 \geq \frac{1}{2} \right\},
\]

\[
\bar{P}^3 = \left\{ \zeta \in \Xi : \zeta_1 \geq 0, \zeta_2 \leq \frac{1}{2} \right\} ; \quad \bar{P}^4 = \left\{ \zeta \in \Xi : \zeta_1 \leq 0, \zeta_2 \leq \frac{1}{2} \right\}.
\]

We readily observe that \( \mathcal{P} \) is a refined partition of \( \mathcal{P} \). The non-anticipatory constraints for the refined partition are given by the set

\[
\mathcal{T}(\mathcal{P}) = \{(1,3,2), (2,4,2)\} \cup \{(k,l)\}_{k,l \in [4], k \leq l}.
\]

Thus, the resulting non-anticipatory constraints are given by \( x_2^1 = x_2^1, x_3^2 = x_3^2 \), and \( x_1^k = x_1^1 \) for all \( k, l \in [4] \). Evaluating the binding constraints for each new decision variable \( x^k \) at the extreme points of each \( \bar{P}^k \), we
conclude that the optimization problem with the refined partition is equivalent to

$$\begin{align*}
\text{minimize} & \quad v \\
\text{subject to} & \quad \max\{-x_2^1, 2x_2^1\} + \max\{x_3^1, 2x_3^1, x_3^2, -2x_3^3\} \leq v \\
& \quad \max\{2x_2^1, 3x_2^1\} + \max\{-\frac{1}{2}x_3^1, -x_3^3\} \leq 0 \\
& \quad \max\{2x_2^1, 3x_2^1\} + \max\{-\frac{1}{2}x_3^1, x_3^3\} \leq 0 \\
& \quad \max\{-x_2^2, 0\} + \max\{2x_2^3, x_3^4\} \leq v \\
& \quad \max\{x_2^2, 2x_2^3\} + \max\{-\frac{1}{2}x_3^2, -x_3^3\} \leq 0 \\
& \quad \max\{x_2^2, 2x_2^3\} + \max\{-\frac{1}{2}x_3^2, x_3^3\} \leq 0 \\
& \quad 1 \leq x_3^4, x_3^4, \frac{1}{2} \leq x_2^3, x_3^3 \leq x_3^3
\end{align*}$$

An optimal solution for this problem is $x_2^3 = x_2^4 = -\frac{2}{3}, x_2^3 = x_2^4 = -1, x_1^3 = x_1^4 = 1, x_2^3 = \frac{1}{2},$ and $x_3^3 = \frac{3}{4},$ with an optimal objective value of $\frac{11}{4} < 4.$ We have thus shown an example where there are no refined partitions which satisfy Condition (a), but the example has a refined partition which satisfies Condition (b) and improves the optimal cost. □

F.3. Partitioning algorithm

We now describe the algorithm for constructing partitions for finite adaptability. The iterative partitioning algorithm is similar to that proposed by Bertsimas and Dunning (2016). At each iteration of the algorithm, we first solve Problem (EC.60) 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.4. The next iteration begins with the refined partition.

In order to describe our heuristic for constructing the refined partition, we introduce some necessary notation. First, for simplicity of exposition, we focus on problems where $d_t = 1$ for all $t = 1, \ldots, T$, though all following results can be readily extended to the general case of $d_t > 1$. Second, we assume that Assumption 4 holds and that $\Xi \subseteq \mathbb{R}^T$ is a hyperrectangle of the form $[\ell, u]$, where the lower bound vector $\ell \in \mathbb{R}^T$ and the upper bound vector $u \in \mathbb{R}^T$ may have components which equal $-\infty$ or $+\infty$, respectively. We define the projection of a vector $\zeta \in \mathbb{R}^d$ onto the first $t$ stages as $\zeta_{1:t} := (\zeta_1, \ldots, \zeta_t) \in \mathbb{R}^t$, and define the projection of a partition $\mathcal{P}$ of $\Xi$ onto its first $t$ stages as $\mathcal{P}_t := \{P^1_t, \ldots, P^K_t\}$.

**Remark EC.2.** The collection $\mathcal{P}_t$ is not necessarily a partition of $\bigcup_{k=1}^K P^k_t$. Nevertheless, we will show that our iterative partitioning scheme always creates partitions such that $\mathcal{P}_t$ is in itself a partition of $\bigcup_{k=1}^K P^k_t$.

We now present the iterative algorithm for constructing partitions for finite adaptability, which is visualized in Figure EC.1. In each iteration of the algorithm, we start with a partition $\mathcal{P}$ of $\Xi$ and solve the corresponding instance of Problem (EC.60). We then construct a refined partition of $\mathcal{P}$ by considering each stage $t = 1, \ldots, T$ sequentially. Specifically, for each stage $t$ and each hyperrectangle $\hat{P} \equiv [\hat{\ell}, \hat{u}] \in \mathcal{P}_t$, we find all of the realizations $\zeta \in \mathcal{A} := \bigcup_{k \in [K]} \mathcal{A}(P^k)$ such that $\zeta_{1:t}$ falls inside $\hat{P}$. We compute the median $M$ of the
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.
Algorithm 1: Iterative Data-driven Partitioning

**Algorithm Partitioning**

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

**Output:** Partition $\mathcal{P} \equiv \{P^1, \ldots, P^K\}$.

Initialize: $K := 1$, $P^1 := \Xi \equiv [t^1, u^1]$, $\mathcal{P} := \{P^1\}$.

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

Solve Problem (EC.60) with partition $\mathcal{P} \equiv \{P^1, \ldots, P^K\}$.

Generate an active set $\mathcal{A} := \bigcup_{k \in [K]} \mathcal{A}(P^k)$.

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

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

done

done

**Procedure AddCuts($\mathcal{P} \equiv \{P^1, \ldots, P^K\}, \mathcal{A}, t$)**

Initialize $\hat{\mathcal{P}} := \mathcal{P}$ and $\hat{K} := K$.

Update $\mathcal{P}_t := \bigcup_{k=1}^{K} \{P^k\}$.

for $\hat{P} \in \mathcal{P}_t$ do

Find $\mathcal{B} := \{\zeta \in \mathcal{A} : \zeta_{1:t} \in \hat{P}\}$.

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

Let $v$ be the ordering of the set $\{\zeta_t : \zeta \in \mathcal{B}\}$.

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 $P \in \mathcal{P}$ such that $P_t = \hat{P}$ do

Set $\hat{P}^1 := P \cap \{\zeta : \zeta_t \leq \text{med}\}$ and $\hat{P}^2 := P \cap \{\zeta : \zeta_t \geq \text{med}\}$.

Update $\hat{\mathcal{P}} := \hat{\mathcal{P}} \cup \{\hat{P}^1, \hat{P}^2\} \setminus \{P\}$ and $\hat{K} := \hat{K} + 1$.

done

done

return $\hat{\mathcal{P}}, \hat{K}$. 
th components of these realizations. Then, for each \( P^k \in \mathcal{P} \) such that \( P^k = \tilde{P} \), we split \( P^k \) along the cut \( \zeta = M \). The entire algorithm is presented formally in Algorithm 1.

Algorithm 1 has several desirable properties as a cut-based recursive partitioning scheme for Problem (EC.60). In particular, Algorithm 1 has a benefit of decoupling the scaling of \( K \) in each iteration from \( N \). Specifically, after the algorithm concludes on iteration \( \text{ITER} \in \mathbb{N} \), we have a guarantee that
\[
K \leq \min\left\{ 2^{\text{ITER} \cdot T}, \log_2(N(1+m)) \right\}.
\]
Thus, the number of partitions formed in each iteration scales lightly in the number of sample path. Additionally, in each iteration of the algorithm and at each stage \( t \in [T] \), we only introduce cuts which satisfy at least one of necessary conditions in Proposition EC.4. Finally, we remark that the algorithm can also be modified to increase the number of allowed cuts per stage to any \( C \geq 2 \), whereby the cuts would be selected as the \( C+1 \)-quantiles of the set \( \{\zeta : \zeta \in B\} \).

Appendix G: Proof of Theorem 5 from Section 6

In this appendix, we present the proof of Theorem 5. We begin with the following intermediary lemma.

**Lemma EC.4.** The \( \infty \)-Wasserstein ambiguity set is equivalent to
\[
\left\{ \frac{1}{N} \sum_{j=1}^{N} Q_j : Q_j \left( \|\xi - \hat{\xi}_j\| \leq \epsilon_N \right) = 1 \quad \forall j \in [N] \right\}.
\]

**Proof.** By the definition of the \( \infty \)-Wasserstein distance from Section 6,
\[
\left\{ Q \in \mathcal{P}(\Xi) : d_\infty \left( Q, \hat{P}_N \right) \leq \epsilon_N \right\} = \left\{ Q \in \mathcal{P}(\Xi) : \right. \begin{array}{l}
\Pi \in \mathcal{P}(\Xi \times \Xi), \\
\Pi \left( \|\xi - \xi'\| \leq \epsilon_N \right) = 1, \text{ and} \\
\Pi \text{ is a joint distribution of } \xi \text{ and } \xi' \\
\text{with marginals } Q \text{ and } \hat{P}_N, \text{ respectively} \end{array} \right\}.
\]

Let \( \tilde{\xi}^1, \ldots, \tilde{\xi}^L \) be the distinct vectors among \( \hat{\xi}^1, \ldots, \hat{\xi}^N \), and let \( I_1, \ldots, I_L \) be index sets defined as
\[
I_t := \{ j \in [N] : \hat{\xi}_j = \tilde{\xi}^t \}.
\]

For any joint distribution \( \Pi \) that satisfies the constraints in the ambiguity set in (EC.66), let \( Q_t \) be the conditional distribution of \( \xi \) given \( \xi' = \tilde{\xi}^t \). Then, for every Borel set \( A \subseteq \Xi \),
\[
Q(\xi \in A) = \Pi ( (\xi, \xi') \in A \times \Xi ) = \sum_{t=1}^{L} \Pi ( \xi \in A \mid \xi' = \tilde{\xi}^t ) \hat{P}_N ( \xi' = \tilde{\xi}^t ) = \sum_{t=1}^{L} Q_t (\xi \in A) \frac{|I_t|}{N}.
\]

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 definitions of \( Q_t \) and \( \hat{P}_N \). Since the above equalities hold for every Borel set, we have shown that
\[
Q = \sum_{t=1}^{L} \frac{|I_t|}{N} Q_t.
\]
Furthermore, by using similar reasoning as above, we observe that
\[ \Pi(\|\xi - \xi'\| \leq \epsilon_N) = \sum_{\ell=1}^L \Pi(\|\xi - \xi'\| = \epsilon_N) \hat{P}_N(\xi' = \hat{\xi}') = \sum_{\ell=1}^L Q_\ell(\|\xi - \hat{\xi}'\| = \epsilon_N) \frac{|I_\ell|}{N}. \]

Combining the above results, the ambiguity set from (EC.66) can be rewritten as
\[
\left\{ \frac{1}{N} \sum_{\ell=1}^L Q_\ell: \sum_{\ell=1}^L Q_\ell(\|\xi - \hat{\xi}'\| \leq \epsilon_N) = 1, \quad Q_1, \ldots, Q_L \in \mathcal{P}(\Xi) \right\} = \left\{ \frac{1}{N} \sum_{j=1}^N Q_j: Q_j(\|\xi - \hat{\xi}'\| \leq \epsilon_N) = 1, \quad Q_1, \ldots, Q_N \in \mathcal{P}(\Xi) \right\}. 
\]

The first equality follows because \( Q_\ell(\|\xi - \hat{\xi}'\| \leq \epsilon_N) = 1 \) for each \( \ell \in [L] \). The second equality follows because \( Q_\ell(\|\xi - \hat{\xi}'\| \leq \epsilon_N) = 1 \) if and only if there exists \( Q_j \in \mathcal{P}(\Xi) \) for each \( j \in I_\ell \) such that \( Q_j(\|\xi - \hat{\xi}'\| \leq \epsilon_N) = 1 \) and \( \sum_{j \in I_\ell} \frac{1}{|I_\ell|} Q_j = Q_\ell \). This concludes the proof. \( \Box \)

We now present the proof of Theorem 5.

**Theorem 5.** Problem (2) with uncertainty sets of the form
\[ \mathcal{U}_N := \left\{ \zeta = (\zeta_1, \ldots, \zeta_T) \in \Xi: \|\zeta - \hat{\xi}'\| \leq \epsilon_N \right\} \]
is equivalent to \( \infty \)-WDRO.

**Proof.** It follows from Lemma EC.4 that the \( \infty \)-Wasserstein ambiguity set can be decomposed into separate distributions, each having a support that is contained in \( \{ \zeta \in \Xi: \|\zeta - \hat{\xi}'\| \leq \epsilon_N \} \) for \( j \in [N] \). Of course, these sets are exactly equal to the uncertainty sets from Section 3, and thus Lemma EC.4 implies that the \( \infty \)-Wasserstein ambiguity set is equivalent to
\[
\left\{ \frac{1}{N} \sum_{j=1}^N Q_j: Q_j \in \mathcal{P}(\mathcal{U}_N) \text{ for each } j \in [N] \right\}.
\]
Therefore, when \( \mathcal{A}_N \) is the \( \infty \)-Wasserstein ambiguity set and each \( \mathcal{U}_N \) is a closed balls around \( \hat{\xi}' \) which is intersected with \( \Xi \),
\[
\sup_{Q \in \mathcal{A}_N} \mathbb{E}_Q \left[ \sum_{t=1}^T c_t(\xi) \cdot x_t(\xi_1, \ldots, \xi_{t-1}) \right] = \frac{1}{N} \sum_{j=1}^N \sup_{Q \in \mathcal{P}(\mathcal{U}_N)} \mathbb{E}_Q \left[ \sum_{t=1}^T c_t(\xi) \cdot x_t(\xi_1, \ldots, \xi_{t-1}) \right] = \frac{1}{N} \sum_{j=1}^N \sum_{\zeta \in \mathcal{U}_N} \sum_{t=1}^T c_t(\zeta) \cdot x_t(\zeta_1, \ldots, \zeta_{t-1}).
\]
Moreover, it similarly follows from Lemma EC.4 that the following inequalities are equivalent:

\[
\mathbb{Q} \left( \sum_{t=1}^{T} \mathbf{A}_t(\xi_t)\mathbf{x}_t(\xi_1, \ldots, \xi_{t-1}) \leq \mathbf{b}(\xi) \right) = 1 \quad \forall \mathbb{Q} \in \mathcal{A}_N \\
\frac{1}{N} \sum_{j=1}^{N} \mathbb{Q}_j \left( \sum_{t=1}^{T} \mathbf{A}_t(\xi_t)\mathbf{x}_t(\xi_1, \ldots, \xi_{t-1}) \leq \mathbf{b}(\xi) \right) = 1 \quad \forall \mathbb{Q}_j \in \mathcal{P}(\mathcal{U}_N^j), \ j \in [N] \\
\mathbb{Q}_j \left( \sum_{t=1}^{T} \mathbf{A}_t(\xi_t)\mathbf{x}_t(\xi_1, \ldots, \xi_{t-1}) \leq \mathbf{b}(\xi) \right) = 1 \quad \forall \mathbb{Q}_j \in \mathcal{P}(\mathcal{U}_N^j), \ j \in [N] \\
\sum_{t=1}^{T} \mathbf{A}_t(\zeta_t)\mathbf{x}_t(\zeta_1, \ldots, \zeta_{t-1}) \leq \mathbf{b}(\zeta) \quad \forall \zeta \in \mathbb{U}_N^j, \ j \in [N].
\]

We have thus shown that Problem (2) and Problem (6) have equivalent objective functions and constraints under the specified constructions of the uncertainty sets and ambiguity set. This concludes the proof. \( \square \)

**Appendix H: Proof of Theorem 6 from Section 6**

In this appendix, we present the proof of Theorem 6.

**Theorem 6.** If \( p \in [1, \infty) \) and \( \epsilon_N > 0 \), then a decision rule is feasible for \( p \)-WDRO only if

\[
\sum_{t=1}^{T} \mathbf{A}_t(\zeta_t)\mathbf{x}_t(\zeta_1, \ldots, \zeta_{t-1}) \leq \mathbf{b}(\zeta) \quad \forall \zeta \in \Xi.
\]

**Proof.** Consider any arbitrary \( \bar{\xi} \in \Xi \) such that \( \bar{\xi} \neq \bar{\xi}^j \) for each \( j \in [N] \). Let \( \delta_{\xi} \) denote the Dirac delta distribution which satisfies \( \delta_{\xi}(\xi = \bar{\xi}) = 1 \), and let \( \hat{\mathbb{P}}_N := \frac{1}{N} \sum_{j=1}^{N} \delta_{\bar{\xi}^j} \) be the empirical distribution of the sample paths. For any \( \lambda \in (0, 1) \), let the convex combination of the two distributions be given by

\[
\mathbb{Q}_\lambda := (1 - \lambda)\hat{\mathbb{P}}_N + \lambda \delta_{\bar{\xi}}.
\]

We recall the definition of the \( p \)-Wasserstein distance between \( \hat{\mathbb{P}}_N \) and \( \mathbb{Q}_\lambda \):

\[
d_p \left( \hat{\mathbb{P}}_N, \mathbb{Q}_\lambda \right) = \inf \left\{ \left( \int_{\Xi \times \Xi} \|\xi - \xi'|^p \ d\Pi(\xi, \xi') \right)^{\frac{1}{p}} : \Pi \text{ is a joint distribution of } \xi \text{ and } \xi' \text{ with marginals } \hat{\mathbb{P}}_N \text{ and } \mathbb{Q}_\lambda, \text{ respectively} \right\}. \quad \text{(EC.67)}
\]

Consider a feasible joint distribution \( \hat{\Pi} \) for the above optimization problem in which \( \xi' \sim \mathbb{Q}_\lambda, \xi'' \sim \hat{\mathbb{P}}_N \), and

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

Indeed, we readily verify that the marginal distributions of \( \xi \) and \( \xi' \) are \( \hat{\mathbb{P}}_N \) and \( \mathbb{Q}_\lambda \), respectively, and thus this joint distribution is feasible for the optimization problem in (EC.67). Moreover,

\[
d_p \left( \hat{\mathbb{P}}_N, \mathbb{Q}_\lambda \right) \leq \left( \int_{\Xi \times \Xi} \|\xi - \xi'|^p \ d\hat{\Pi}(\xi, \xi') \right)^{\frac{1}{p}}
\]
\[
\left( \int_{\Xi \times \Xi} \| \xi - \xi' \|^p \mathbb{I} \{ \xi' = \bar{\xi} \} \, d\Pi(\xi, \xi') + \int_{\Xi \times \Xi} \| \xi - \xi' \|^p \mathbb{I} \{ \xi' \neq \bar{\xi} \} \, d\Pi(\xi, \xi') \right)^{\frac{1}{p}} = \left( \frac{1}{N} \sum_{j=1}^{N} \lambda \| \hat{\xi}^j - \bar{\xi} \|^p \right)^{\frac{1}{p}}.
\]

The inequality follows since \( \bar{\Pi} \) is a feasible but possibly suboptimal joint distribution for the optimization problem in (EC.67). The first equality follows from splitting the integral into two cases, and observing that the second case equals zero since \( \xi = \xi' \) whenever \( \xi' = \bar{\xi} \). The final equality follows because \( \xi = \xi'' \) whenever \( \xi' = \bar{\xi} \), and \( \xi'' \) is distributed uniformly over the historical sample paths. Thus, for any arbitrary choice of \( \bar{\xi} \in \Xi \), we have shown that \( Q^\lambda_{\bar{\xi}} \) is contained in the \( p \)-Wasserstein ambiguity set whenever \( \lambda \in (0, 1) \) satisfies

\[
\left( \frac{1}{N} \sum_{j=1}^{N} \lambda \| \hat{\xi}^j - \bar{\xi} \|^p \right)^{\frac{1}{p}} \leq \epsilon_N
\]

\[
\frac{1}{N} \sum_{j=1}^{N} \lambda \| \hat{\xi}^j - \bar{\xi} \|^p \leq \epsilon_N^p
\]

\[
\lambda \leq \frac{\epsilon_N^p}{\frac{1}{N} \sum_{j=1}^{N} \| \hat{\xi}^j - \bar{\xi} \|^p}.
\]

Now, consider any feasible decision rule for Problem (6), i.e., a decision rule \( x \in \mathcal{X} \) which satisfies

\[
\sum_{t=1}^{T} A_t(\xi) x_t(\xi_1, \ldots, \xi_{t-1}) \leq b(\xi) \quad \text{Q-a.s., } \forall \xi \in \mathcal{A}_N.
\]  

(EC.68)

Let \( \mathcal{A}_N \) be the \( p \)-Wasserstein ambiguity set for \( 1 \leq p < \infty \) and \( \epsilon_N > 0 \). Then, for any arbitrary \( \bar{\xi} \in \Xi \), there exists a \( \lambda \in (0, 1) \) such that \( Q_{\bar{\xi}}^\lambda \) is contained in \( \mathcal{A}_N \), and so it follows from (EC.68) that the decision rule must satisfy

\[
\sum_{t=1}^{T} A_t(\bar{\xi}) x_t(\bar{\xi}_1, \ldots, \bar{\xi}_{t-1}) \leq b(\bar{\xi}).
\]

Since \( \bar{\xi} \in \Xi \) was chosen arbitrarily, we conclude that the decision rule must satisfy

\[
\sum_{t=1}^{T} A_t(\xi) x_t(\xi_1, \ldots, \xi_{t-1}) \leq b(\xi) \quad \forall \xi \in \Xi,
\]

which is what we wished to show. \[ \square \]

**Appendix I: Finite Adaptability for Problem (6) with 1-Wasserstein Ambiguity Sets**

In this appendix, we present a reformulation of finite adaptability for Problem (6) using the 1-Wasserstein ambiguity set. Specifically, we present a variant on the reformulation of finite adaptability with linear decision rules from Section 5.2 in the context of Problem (6) with the 1-Wasserstein ambiguity set described in Section 6. The performance of this data-driven approach is illustrated in Section 7.
We first review the necessary notation. Following Section 5, we focus on a specific case of Problem (6) of the form

\[
\begin{aligned}
\min_{x \in \mathcal{X}} & \quad \sup_{Q \in \mathcal{A}_N} \mathbb{E}_Q \left[ \sum_{t=1}^{T} c_t \cdot x_t(\xi_1, \ldots, \xi_{t-1}) \right] \\
\text{subject to} & \quad \sum_{t=1}^{T} A_t x_t(\xi_1, \ldots, \xi_{t-1}) \leq b(\xi) \quad \forall Q \in \mathcal{A}_N,
\end{aligned}
\]  

(EC.69)

in which \( A_t(\xi) \) and \( c_t(\xi) \) do not depend on the stochastic process. The ambiguity set is constructed as

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

where \( \hat{\mathcal{P}}_N \) is the empirical distribution of the historical data, \( \epsilon_N \geq 0 \) is a robustness parameter, and the 1-Wasserstein distance between two distributions is given by

\[
d_1(Q, Q') = \inf \left\{ \int_{\Xi \times \Xi} \| \xi - \xi' \| d\Pi(\xi, \xi') : \Pi \text{ is a joint distribution of } \xi \text{ and } \xi', \text{ with marginals } Q \text{ and } Q', \text{ respectively} \right\}.
\]

We refer to Section 6 for more details on the 1-Wasserstein ambiguity set. We assume that the robustness parameter satisfies \( \epsilon_N > 0 \), in which case it follows from Theorem 6 in Section 6 that Problem (EC.69) is equivalent to

\[
\begin{aligned}
\min_{x \in \mathcal{X}} & \quad \sup_{Q \in \mathcal{A}_N} \mathbb{E}_Q \left[ \sum_{t=1}^{T} c_t \cdot x_t(\xi_1, \ldots, \xi_{t-1}) \right] \\
\text{subject to} & \quad \sum_{t=1}^{T} A_t x_t(\xi_1, \ldots, \xi_{t-1}) \leq b(\xi) \quad \forall \xi \in \Xi.
\end{aligned}
\]  

(EC.70)

We next present an extension of finite adaptability to Problem (EC.70). Let \( P^1, \ldots, P^K \) be a given partition of \( \Xi \subseteq \mathbb{R}^d \). Then, we consider restricting the space of decision rules to those of the form

\[
x_t(\xi_1, \ldots, \xi_{t-1}) = \begin{cases} x^1_{t,0} + \sum_{s=1}^{t-1} X^1_{t,s} \xi_s, & \text{if } (\xi_1, \ldots, \xi_{t-1}) \in P^1_{t-1}, \\ \vdots & \\ x^K_{t,0} + \sum_{s=1}^{t-1} X^K_{t,s} \xi_s, & \text{if } (\xi_1, \ldots, \xi_{t-1}) \in P^K_{t-1}, \end{cases}
\]

where \( P^k_{t-1} \) denotes the projection of \( P^k \) onto the first \( t \) time stages. The resulting approximation of Problem (EC.70) is given by

\[
\begin{aligned}
\min_{x \in \mathcal{X}} & \quad \sup_{Q \in \mathcal{A}_N} \mathbb{E}_Q \left[ \max_{k \in [K]} \sum_{\xi \in P^k} \sum_{t=1}^{T} c_t \cdot \left( x^k_{t,0} + \sum_{s=1}^{t-1} X^k_{t,s} \xi_s \right) \right] \\
\text{subject to} & \quad \sum_{t=1}^{T} A_t \left( x^k_{t,0} + \sum_{s=1}^{t-1} X^k_{t,s} \xi_s \right) \leq b(\xi) \quad \forall \xi \in P^k, \quad k \in [K] \\
x^k_{t,0} = x^k_{t,t}, \quad X^k_{t,s} = X^k_{t,s} & \quad \forall (k, k', t) \in T(P^1, \ldots, P^K), \quad 1 \leq s < t,
\end{aligned}
\]  

(EC.71)

where the decision variables are \( x^k_{t,0} \in \mathbb{R}^{n_t} \) and \( X^k_{t,s} \in \mathbb{R}^{n_t \times d_s} \) for all \( 1 \leq s < t \) and \( k \). For further details about finite adaptability, we refer to Section 5.2.

In the remainder of this appendix, we develop a tractable reformulation of Problem (EC.71). Our reformulation, which will use similar duality techniques to those presented in Section 5, is presented as Theorem EC.3 at the end of this appendix. Our reformulation requires the following assumption:
Assumption EC.1. The partition $P^1, \ldots, P^K$ of $\Xi \subseteq \mathbb{R}^d$ has the form $P^k := \{ \zeta \in \mathbb{R}^d : \ell^k \leq \zeta \leq u^k \} \neq \emptyset$ for each $k \in [K]$ ($\ell^k, u^k \in \mathbb{R}^d$). Moreover, the norm in the 1-Wasserstein distance is equal to $\| \cdot \|_1$.

We remark that the conditions on the partition is a mild assumption that is satisfied, for example, when $\Xi = \mathbb{R}^d_+$ is the nonnegative orthant and the partitions are selected using the algorithm from Appendix F.

Before presenting the reformulation, we first introduce two useful intermediary results. The first intermediary result (Lemma EC.5) uses reformulation techniques for the 1-Wasserstein ambiguity set from Esfahani and Kuhn (2018) to reformulate the objective function in Problem (EC.71).

**Lemma EC.5.** If Assumption EC.1 holds, then Problem (EC.71) is equivalent to

\[
\begin{aligned}
&\text{minimize } \lambda \epsilon_N + \frac{1}{N} \sum_{j=1}^N v_j \\
&\text{subject to } \sum_{t=1}^T c_t \left( x_{t,0}^k + \sum_{s=1}^{t-1} x_{t,s}^k \zeta_s \right) - \lambda \| \zeta - \hat{\zeta}^t \|_1 \leq v_j \quad \forall \zeta \in P^k, j \in [N], k \in [K] \\
&\quad \sum_{t=1}^T A_t \left( x_{t,0}^k + \sum_{s=1}^{t-1} x_{t,s}^k \zeta_s \right) \leq b(\zeta) \quad \forall \zeta \in P^k, k \in [K] \\
&\quad x_{t,0}^k = x_{t,0}^{k'}, x_{t,s}^k = x_{t,s}^{k'}, \quad (k, k', t) \in T(P^1, \ldots, P^K), 1 \leq s < t,
\end{aligned}
\]

(EC.72)

where the decision variables are $x_{t,0}^k \in \mathbb{R}^n$ and $x_{t,s}^k \in \mathbb{R}^{n_t \times d_s}$ for all $1 \leq s < t$ and $k, v \in \mathbb{R}^N$, and $\lambda \in \mathbb{R}_+$.

**Proof.** For any fixed solution to Problem (EC.71), let

\[
\ell_k(\zeta) := \begin{cases} 
\sum_{t=1}^T c_t \left( x_{t,0}^k + \sum_{s=1}^{t-1} x_{t,s}^k \zeta_s \right), & \text{if } \zeta \in P^k, \\
-\infty, & \text{otherwise}
\end{cases}
\]

for each $k \in [K]$ and $\zeta \in \Xi$. It follows from Assumption EC.1 that $\Xi \subseteq \mathbb{R}^d$ and the regions $P^1, \ldots, P^K$ are nonempty, convex, and closed. Therefore, we observe that the set $\Xi$ and the functions $\ell_k$ for each $k \in [K]$ satisfy Esfahani and Kuhn (2018, Assumption 4.1). Therefore,

\[
\begin{aligned}
&\sup_{Q \in A_N} \mathbb{E}_Q \left[ \max_{k \in [K]} \sum_{t=1}^T c_t \left( x_{t,0}^k + \sum_{s=1}^{t-1} x_{t,s}^k \zeta_s \right) \right] = \sup_{Q \in A_N} \mathbb{E}_Q \left[ \max_{k \in [K]} \ell_k(\zeta) \right] \\
&= \inf_{\lambda \geq 0} \lambda \epsilon_N + \frac{1}{N} \sum_{j=1}^N \sup_{\zeta \in \Xi} \left( \max_{k \in [K]} \ell_k(\zeta) - \lambda \| \zeta - \hat{\zeta}^t \|_1 \right) \\
&= \inf_{\lambda \geq 0} \lambda \epsilon_N + \frac{1}{N} \sum_{j=1}^N \max_{k \in [K]} \sup_{\zeta \in \Xi} \left( \ell_k(\zeta) - \lambda \| \zeta - \hat{\zeta}^t \|_1 \right).
\end{aligned}
\]

The first equality follows from the definition of $\ell_k$ for each $k \in [K]$. The second equality follows from Esfahani and Kuhn (2018, Equation 12b). The final equality follows because $P^1, \ldots, P^K$ form a partition of $\Xi$. Adding auxiliary epigraph decision variables $v \in \mathbb{R}^N$, we obtain the desired reformulation of Problem (EC.71).

Our second intermediary result characterizes optimal solutions for a class of linear optimization problems.

**Lemma EC.6.** Consider the optimization problem

\[
\begin{aligned}
&\text{minimize } \alpha a + \beta b \\
&\text{subject to } |d - \alpha + \beta| \leq \lambda,
\end{aligned}
\]

(EC.73)
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. We observe that \( \lambda \geq 0 \) implies that Problem (EC.73) has a feasible solution, and \( a + b \geq 0 \) implies that Problem (EC.73) has a finite optimal cost. Moreover, Problem (EC.73) is equivalent to the following linear optimization problem:

\[
\begin{align*}
\text{minimize} \quad & \alpha a + \beta b \\
\text{subject to} \quad & d - \alpha + \beta \leq \lambda, \\
& d - \alpha + \beta \geq -\lambda.
\end{align*}
\] (EC.74)

Furthermore, since \( \alpha, \beta \in \mathbb{R}_+ \), it follows from Bertsimas and Tsitsiklis (1997, Theorem 2.6) that Problem (EC.74) has at least one extreme point, and thus has an optimal solution that is an extreme point (Bertsimas and Tsitsiklis 1997, Theorem 2.7). We observe that each extreme point of Problem (EC.74) satisfies \( \alpha = 0 \) or \( \beta = 0 \). From this point, we focus on optimal solutions to Problem (EC.73) which are optimal extreme points for Problem (EC.74).

Suppose \( a > 0 \) and \( b \leq 0 \). We have two cases to consider.

- If \( d > \lambda \), then any optimal extreme point \((\alpha^*, \beta^*)\) must satisfy \( \alpha^* \geq d - \lambda > 0 \). Since \( \alpha^* \) and \( \beta^* \) cannot both be positive at an extreme point, we conclude that \( \beta^* = 0 \) and \( \alpha^* = d - \lambda \).

- If \( d \leq \lambda \), then \((0, 0)\) is a feasible extreme point. Moreover, any extreme point of the form \((\alpha, 0)\) for \( \alpha > 0 \) will produce a worse objective value compared to \((0, 0)\). Thus, any optimal extreme point \((\alpha^*, \beta^*)\) must satisfy \( \alpha^* = 0 \). Since \( b \leq 0 \), an optimal extreme point is given by \( \alpha^* = 0 \) and \( \beta^* = \lambda - d \).

Thus, we have shown that \( \alpha^* = [d - \lambda]_+ \) and \( \beta^* = [\lambda - d]_+ \) is an optimal solution. The case in which \( a \leq 0 \) and \( b > 0 \) follows by identical reasoning.

Suppose \( a \geq 0 \) and \( b \geq 0 \). Then, it follows from the constraints of Problem (EC.74) and nonnegativity of \( \alpha \) and \( \beta \) that any feasible solution to Problem (EC.74) must satisfy \( \alpha \geq \max\{d - \lambda, 0\} \) and \( \beta \geq \max\{-\lambda - d, 0\} \). Since \( a \) and \( b \) are both nonnegative, an optimal extreme point has the form \( \alpha^* = [d - \lambda]_+ \) and \( \beta^* = [-d - \lambda]_+ \).

Note that both \( \alpha^* \) and \( \beta^* \) may equal their minimal feasible value while remaining an extreme point of the problem.

Since \( a + b \geq 0 \), we do not need to treat the case where \( a < 0 \) and \( b < 0 \). Combining the different conditions for \( a \) and \( b \), we obtain the desired result. \( \square \)

We now combine the above lemmas with techniques from Section 5 to establish our reformulation.

Theorem EC.3. If Assumption EC.1 holds, then Problem (EC.72) can be reformulated by adding at most \( O(Kmd) \) additional continuous decision variables and \( O(NK + Kmd) \) additional linear constraints. Defining
the reformulation is given by

\[
\text{minimize } \lambda \epsilon_N + \frac{1}{N} \sum_{j=1}^{N} v_j \\
\text{subject to } \sum_{t=1}^{\ell_t} \mathbf{c}_t \cdot \left( \mathbf{x}_{t,0}^k + \sum_{s=1}^{t-1} \mathbf{x}_{t,s}^k \right) + \sum_{p=1}^{3} \sum_{i \in \mathcal{I}_i^k} \alpha_i^{kp} (u_i^t - \tilde{\xi}_i^t) + \beta_i^{kp} (\tilde{\xi}_i^t - \ell_i^t) \leq v_j \quad j \in [N], \ k \in [K],
\]

\[
\left\| \sum_{s=t+1}^{T} (\mathbf{X}_{t,s}^k)^T \mathbf{c}_s - \alpha^{kp} + \beta^{kp} \right\|_{\infty} \leq \lambda \quad p \in [3], \ k \in [K], \ t \in [T],
\]

\[
\sum_{t=1}^{T} (\mathbf{M}_t^u u_t^k - \Lambda_t^k \ell_t^k + \mathbf{A}_t \mathbf{x}_{t,0}^k) \leq b^0 \quad k \in [K],
\]

\[
\mathbf{M}_t^k - \Lambda_t^k = -\mathbf{B}_t + \sum_{a=t+1}^{T} \mathbf{A}_a \mathbf{x}_{a,t}^k \quad k \in [K], \ t \in [T],
\]

\[
\mathbf{x}_{t,0}^k = \mathbf{x}_{t,0}^{k',}\mathbf{x}_{t,s}^k = \mathbf{x}_{t,a}^{k'} \quad (k,k',t) \in \mathcal{T}(\mathcal{S}), \ 1 \leq s < t.
\]

where the auxiliary decision variables are \( \alpha^{kp} := (\alpha_1^{kp}, \ldots, \alpha_3^{kp}), \beta^{kp} := (\beta_1^{kp}, \ldots, \beta_3^{kp}) \in \mathbb{R}^d \) for all \( k \in [K] \) and \( p \in [3] \), as well as \( \mathbf{M}^k := (\mathbf{M}_1^k, \ldots, \mathbf{M}_T^k), \Lambda^k := (\Lambda_1^k, \ldots, \Lambda_T^k) \in \mathbb{R}^{m \times d} \) for all \( k \in [K] \).

### Proof of Theorem EC.3.

Choose any \( k \in [K] \). Following similar reasoning to Theorem 3, the constraints

\[
\sum_{t=1}^{T} \mathbf{A}_t \left( \mathbf{x}_{t,0}^k + \sum_{s=1}^{t-1} \mathbf{x}_{t,s}^k \right) \leq b^0 + \sum_{t=1}^{T} \mathbf{B}_t \zeta_t \quad \forall \zeta \in \mathbb{P}^k,
\]

are satisfied if and only if there exist \( \mathbf{M}^k := (\mathbf{M}_1^k, \ldots, \mathbf{M}_T^k), \Lambda^k := (\Lambda_1^k, \ldots, \Lambda_T^k) \in \mathbb{R}^{m \times d} \) which satisfy

\[
\sum_{t=1}^{T} (\mathbf{M}_t u_t^k - \Lambda_t^k \ell_t^k + \mathbf{A}_t x_{t,0}^k) \leq b^0,
\]

\[
\mathbf{M}_t^k - \Lambda_t^k = \sum_{a=t+1}^{T} \mathbf{A}_a x_{a,t}^k - \mathbf{B}_t, \quad t \in [T].
\]

The remainder of the proof focuses on the epigraph constraints. Recall from Assumption EC.1 that \( \mathbb{P}^k \) 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.72) is equivalent to

\[
\sup_{\mathbf{a} \leq \mathbf{c} \leq \mathbf{u}} \sum_{t=1}^{T} \mathbf{c}_t \cdot \left( \mathbf{x}_{t,0}^k + \sum_{s=1}^{t-1} \mathbf{x}_{t,s}^k \right) - \lambda \sum_{t=1}^{T} \| \zeta_t - \tilde{\zeta}_t \|_1 \leq v_j. \quad \text{(EC.75)}
\]

It follows from strong duality of linear programming that the left-hand side of (EC.75) is equivalent to

\[
\text{minimize } \mathbf{a}, \mathbf{b} \in \mathbb{R}^d \sum_{t=1}^{T} \left( \mathbf{c}_t \cdot \mathbf{x}_{t,0}^k + \mathbf{a}_t \cdot (u_t^k - \tilde{\xi}_t^k) + \mathbf{b}_t \cdot (\tilde{\xi}_t^k - \ell_t^k) \right) \quad \text{subject to } \left\| \sum_{s=t+1}^{T} (\mathbf{X}_{t,s}^k)^T \mathbf{c}_s - \mathbf{a}_t + \mathbf{b}_t \right\|_{\infty} \leq \lambda, \quad t \in [T]. \quad \text{(EC.76)}
\]
Thus, for each $j$, the constraints in Problem (EC.75) are satisfied if and only if there exists $\alpha_j$ (alternatively, $\beta_j$) such that $\alpha_j(u_j^k - \hat{\xi}_j^k) + \beta_j(\hat{\xi}_j^k - \ell_j^k)$ should be set to zero and the term $\alpha_j(u_j^k - \hat{\xi}_j^k)$ (alternatively, $\beta_j(\hat{\xi}_j^k - \ell_j^k)$) should be dropped from the objective.

We observe that the optimal cost of Problem (EC.76) is equal to the sum over the optimal costs of $d$ separate optimization problems, each of the form

$$\min_{\alpha_t, \beta_t \in \mathbb{R}^+} \alpha_t (u_t^k - \hat{\xi}_t^k) + \beta_t (\hat{\xi}_t^k - \ell_t^k)$$

subject to $|g_t^k - \alpha_t + \beta_t| \leq \lambda$.

where $g^k := (\sum_{s=2}^{T} (X_{s,1}^k)^\top c_s, \sum_{s=3}^{T} (X_{s,2}^k)^\top c_s, \ldots, (X_{S,T}^k)^\top c_T, 0) \in \mathbb{R}^d$. Since $(u_t^k - \hat{\xi}_t^k) - (\ell_t^k - \hat{\xi}_t^k) \geq 0$, Lemma EC.6 implies that the optimal value of $\alpha_t$ and $\beta_t$ can be determined by checking whether $\hat{\xi}_t^k < \ell_t^k$, $\hat{\xi}_t^k > u_t^k$, or $\ell_t^k \leq \hat{\xi}_t^k \leq u_t^k$. This information is captured by the index sets

$$\mathcal{I}^{kj}_1 = \{l \in [d] : \hat{\xi}_t^k < \ell_t^k\},$$
$$\mathcal{I}^{kj}_2 = \{l \in [d] : \hat{\xi}_t^k > u_t^k\},$$
$$\mathcal{I}^{kj}_3 = \{l \in [d] : \ell_t^k \leq \hat{\xi}_t^k \leq u_t^k\}.$$

Thus, for each $j \in [N]$ and $k \in [K]$, Problem (EC.76) is equivalent to

$$\min_{\alpha_p, \beta_p \in \mathbb{R}^+} \sum_{t=1}^{T} c_t^k \cdot x_{t,0}^k + \sum_{p=1}^{3} \sum_{l \in \mathcal{I}^{kj}_p} \alpha_p (u_t^k - \hat{\xi}_t^k) + \beta_p (\hat{\xi}_t^k - \ell_t^k)$$

subject to $\left\| \sum_{s=t+1}^{T} (X_{s,t}^k)^\top c_s - \alpha_p + \beta_p \right\|_\infty \leq \lambda$,

$$p \in [3], t \in [T].$$

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 Problem (EC.75) are satisfied if and only if there exists $\alpha^{pk} \in \mathbb{R}^d_+$ and $\beta^{pk} \in \mathbb{R}^d_+$ for each $k \in [K]$ and $p \in [3]$ which satisfy

$$\sum_{t=1}^{T} c_t \cdot x_{t,0}^k + \sum_{p=1}^{3} \sum_{l \in \mathcal{I}^{kj}_p} \alpha^{kp} (u_t^k - \hat{\xi}_t^k) + \beta^{kp} (\hat{\xi}_t^k - \ell_t^k) \leq v_j, \quad j \in [N], k \in [K],$$

$$\left\| \sum_{s=t+1}^{T} (X_{s,t}^k)^\top c_s - \alpha^{kp} + \beta^{kp} \right\|_\infty \leq \lambda, \quad p \in [3], k \in [K], t \in [T].$$

Plugging these reformulated constraints back into Problem (EC.72), we obtain the desired reformulation of the epigraph constraints.