A data-driven approach to multi-stage stochastic 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
Department of Information and Decision Sciences, University of Illinois at Chicago, bsturt@uic.edu

We propose a new data-driven approach for addressing multi-stage stochastic linear optimization problems with unknown distributions. The approach consists of solving a robust optimization problem that is constructed from sample paths of the underlying stochastic process. We provide asymptotic bounds on the gap between the optimal costs of the robust optimization problem and the underlying stochastic problem as more sample paths are obtained, and we characterize cases in which this gap is equal to zero. To the best of our knowledge, this is the first sample-path approach for multi-stage stochastic linear optimization that offers asymptotic optimality guarantees when uncertainty is arbitrarily correlated across time. Finally, we develop approximation algorithms for the proposed approach by extending techniques from the robust optimization literature, and demonstrate their practical value through numerical experiments on stylized data-driven inventory management problems.

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

History: First submitted on Nov. 3, 2018. Revisions submitted on April 5, 2020 and April 16, 2021. Accepted for publication on August 22, 2021.

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 stochastic 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 optimization community (Birge and Louveaux 2011, Shapiro et al. 2009).

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 demand trajectories 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 considered 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. Moreover, in the absence of additional assumptions on the estimated distribution or on the problem setting, multi-stage stochastic linear optimization problems are notorious for being computationally demanding.

The main contribution of this paper is a new data-driven approach for multi-stage stochastic linear optimization that can be asymptotically optimal even when uncertainty is arbitrarily correlated across time. In other words, we propose a data-driven approach for addressing multi-stage stochastic linear optimization with unknown distributions that (i) does not require any parametric modeling assumptions on the correlation structure of the underlying probability distribution, and (ii) converges to the underlying multi-stage stochastic linear optimization problem under certain conditions as the size of the dataset tends to infinity. Such asymptotic optimality guarantees are of practical importance, as they ensure that the approach offers a near-optimal approximation of the underlying stochastic problem in the presence of big data.

Our approach for multi-stage stochastic linear optimization is based on robust optimization. Specifically, given sample paths of the underlying stochastic process, the proposed approach consists of constructing and solving a multi-stage robust linear optimization problem with multiple uncertainty sets. The main result of this paper (Theorem 1) establishes, under certain assumptions, that the optimal cost of this robust optimization problem converges nearly to that of the stochastic problem as the number of sample paths tends to infinity. While this robust optimization problem is computationally demanding to solve exactly, we provide evidence that it can be tractably approximated to reasonable accuracy by leveraging approximation techniques from the robust optimization literature. To the best of our knowledge, there was no similar work in the literature which addresses multi-stage stochastic linear optimization by solving a sequence of robust optimization problems.

The paper is organized as follows. Section 2 introduces multi-stage stochastic linear optimization in a data-driven setting. Section 3 presents the new data-driven approach to multi-stage stochastic linear optimization. Section 4 states the main asymptotic optimality guarantees. Section 5 presents two examples of approximation algorithms by leveraging techniques from robust optimization. Section 6 discusses implications of our asymptotic optimality guarantees in the context of Wasserstein-based distributionally robust optimization. Sections 7 and 8 demonstrate the practical value of the proposed methodologies in computational experiments. Section 9 offers concluding thoughts. All technical proofs are relegated to the attached appendices.

1.1. Related literature

Originating with Soyster (1973) and Ben-Tal and Nemirovski (1999), robust optimization has been widely studied as a general framework for decision-making under uncertainty, in which "optimal" decisions are those which perform best under the worst-case parameter realization from an "uncertainty set". Beginning with the seminal work of Ben-Tal et al. (2004), robust optimization has been viewed with particular success as a computationally tractable framework for addressing multi-stage problems. Indeed, by restricting the space of decision rules, a stream of literature showed that multi-stage robust linear optimization problems can be solved in polynomial time by using duality-based reformulations or cutting-plane methods. For a modern overview of decision-rule approximations, we refer the reader to Delage and Iancu (2015), Georghiou et al. (2018), Ben-Tal et al. (2009), Bertsimas et al. (2011a). A variety of non-decision rule approaches to solving multi-stage robust optimization have been proposed as well, such as Zeng and Zhao (2013), Zhen et al. (2018), Xu and Burer (2018), Georghiou et al. (2019).

Despite its computational tractability for multi-stage problems, a central critique of traditional robust optimization is that it does not aspire to find solutions which perform well on average. Several works have aimed to quantify the quality of solutions from multi-stage robust linear optimization from the perspective of multi-stage stochastic linear optimization (Chen et al. 2007, Bertsimas and Goyal 2010, Bertsimas et al. 2011b). By and large, it is fair to say that multi-stage robust linear optimization is viewed today as a distinct framework from multi-stage stochastic linear optimization, aiming to find solutions with good worst-case as opposed to good average performance.

Providing a potential tradeoff between the stochastic and robust frameworks, 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. (2019) for multi-stage distributionally robust linear optimization with moment-based ambiguity sets.

There has been a proliferation of data-driven constructions of ambiguity sets which offer various probabilistic performance guarantees, including those based on the p-Wasserstein distance for $p \in [1,\infty)$ (Pflug and Wozabal 2007, Mohajerin Esfahani and Kuhn 2018), phi-divergences (Ben-Tal et al. 2013, Bayraksan and Love 2015, Van Parys et al. 2020), and statistical hypothesis tests (Bertsimas et al. 2018). 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 three or more stages) directly from data, retain their asymptotic optimality guarantees.

In contrast to the above literature, our motivation for robust optimization in this paper is not to find solutions which perform well on the worst-case realization in an uncertainty set, are risk-averse, or have finite-sample probabilistic guarantees. Rather, our proposed approach to multi-stage stochastic linear optimization adds robustness to the historical data as a tool to avoid overfitting as the number of data points tends to infinity. In this spirit, our work is perhaps closest related to several papers in the context of machine learning (Xu et al. 2012, Shafieezadeh-Abadeh et al. 2019), which showed that adding robustness to historical data can be used to develop machine learning methods which have nonparametric performance guarantees when the solution space (of classification or regression models) is not finite-dimensional. To the best of our knowledge, this paper is the first to apply this use of robust optimization in the context of multi-stage stochastic linear optimization to achieve asymptotic optimality without restricting the space of decision rules.

As far as we are aware, our data-driven approach of averaging over multiple uncertainty sets is novel in the context of multi-stage stochastic linear optimization, and its asymptotic optimality guarantees do not follow from existing literature. 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). In this paper, we instead establish general asymptotic optimality guarantees for the proposed data-driven approach for multi-stage stochastic linear optimization by developing new bounds for distributionally robust optimization with the 1-Wasserstein ambiguity set and connections with nonparametric support estimation (Devroye and Wise 1980).

Under a particular construction of the uncertainty sets, we show that the proposed data-driven approach to multi-stage stochastic linear optimization can also be interpreted as distributionally robust optimization using the ∞-Wasserstein ambiguity set (see Section 6). However, the asymptotic optimality guarantees in our paper do not make use of this interpretation, as there were surprisingly few previous convergence results for this ambiguity set, even in single-stage settings. Indeed, when an underlying distribution is unbounded, the ∞-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 ∞-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 Mohajerin Esfahani and Kuhn (2018, Theorem 3.6) to establish convergence guarantees for the 1-Wasserstein ambiguity set do not appear to extend to the ∞-Wasserstein ambiguity set. As a byproduct of the results in this paper, we obtain asymptotic optimality guarantees for distributionally robust optimization with the ∞-Wasserstein ambiguity set under the same mild probabilistic assumptions as Mohajerin Esfahani and Kuhn (2018) for the first time.

1.2. Notation

We denote the real numbers by $\mathbb{R}$, the nonnegative real numbers by $\mathbb{R}_+$, and the integers by $\mathbb{Z}$. Lowercase and uppercase bold letters refer to vectors and matrices. We assume throughout that $\|\cdot\|$ refers to an ℓ_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, $\text{int}(\cdot)$ be the interior of a set, and $[K]$ be shorthand for the set of consecutive integers $\{1,\dots,K\}$. Throughout the paper, we let $\boldsymbol{\xi} \coloneqq (\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_T) \in \mathbb{R}^d$ denote a stochastic process with a joint probability distribution $\mathbb{P}$, and assume that $\hat{\boldsymbol{\xi}}^1,\dots,\hat{\boldsymbol{\xi}}^N$ are independent and identically distributed (i.i.d.) samples from that distribution. Let $\mathbb{P}^N \coloneqq \mathbb{P}\times\cdots\times\mathbb{P}$ denote the N-fold probability distribution over the historical data. We let $S \subseteq \mathbb{R}^d$ denote the support of $\mathbb{P}$, that is, the smallest closed set where $\mathbb{P}(\boldsymbol{\xi}\in S) = 1$. The extended real numbers are defined as $\bar{\mathbb{R}} \coloneqq \mathbb{R}\cup\{-\infty,\infty\}$, and we adopt the convention that $\infty - \infty = \infty$. The expectation of a measurable function $f:\mathbb{R}^d \to \mathbb{R}$ applied to the stochastic process is denoted by $\mathbb{E}[f(\boldsymbol{\xi})] = \mathbb{E}_{\mathbb{P}}[f(\boldsymbol{\xi})] = \mathbb{E}_{\mathbb{P}}[\max\{f(\boldsymbol{\xi}),0\}] - \mathbb{E}_{\mathbb{P}}[\max\{-f(\boldsymbol{\xi}),0\}]$. 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 $\mathbb{Q}(\boldsymbol{\xi}\in\mathcal{Z}) \equiv \mathbb{E}_{\mathbb{Q}}[\mathbb{I}\{\boldsymbol{\xi}\in\mathcal{Z}\}] = 1$.

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 $\boldsymbol{\xi} \coloneqq (\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_T)\in\mathbb{R}^d$ with an underlying joint probability distribution, where $\boldsymbol{\xi}_t \in \mathbb{R}^{d_t}$ is a random variable that is observed immediately after the decision in stage t is selected. We assume throughout that the random variables $\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_T$ may be correlated. A decision rule $\mathbf{x} \coloneqq (\mathbf{x}_1,\dots,\mathbf{x}_T)$ 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, a policy in each stage is a measurable function of the form $\mathbf{x}_t : \mathbb{R}^{d_1}\times\cdots\times\mathbb{R}^{d_{t-1}} \to \mathbb{R}^{n_t - p_t}\times\mathbb{Z}^{p_t}$. We use the shorthand notation $\mathcal{X}$ to denote the space of all 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

$$\begin{aligned} \underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \quad & \mathbb{E}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi})\cdot\mathbf{x}_t(\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_{t-1})\right] \\ \text{subject to} \quad & \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\xi})\mathbf{x}_t(\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_{t-1}) \le \mathbf{b}(\boldsymbol{\xi}) \quad \text{a.s.} \end{aligned} \tag{1}$$

Following standard convention, we assume that the problem parameters $\mathbf{c}_1(\boldsymbol{\xi}) \in \mathbb{R}^{n_1}, \ldots, \mathbf{c}_T(\boldsymbol{\xi}) \in \mathbb{R}^{n_T}$, $\mathbf{A}_1(\boldsymbol{\xi}) \in \mathbb{R}^{m \times n_1}, \ldots, \mathbf{A}_T(\boldsymbol{\xi}) \in \mathbb{R}^{m \times n_T}$, and $\mathbf{b}(\boldsymbol{\xi}) \in \mathbb{R}^m$ are affine functions of the stochastic process.

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

$$\hat{\boldsymbol{\xi}}^j \equiv (\hat{\boldsymbol{\xi}}_1^j, \ldots, \hat{\boldsymbol{\xi}}_T^j), \qquad j = 1, \ldots, N.$$

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. We assume that $\hat{\boldsymbol{\xi}}^1, \ldots, \hat{\boldsymbol{\xi}}^N$ are independent and identically distributed (i.i.d.) realizations of the stochastic process $\boldsymbol{\xi} \equiv (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_T)$. Our goal in this paper is a general-purpose, nonparametric sample-path approach for solving Problem (1) in practical computation times.

We will also assume that the support of the stochastic process 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, $\mathbb{P}(\boldsymbol{\xi} \in \Xi) = 1$.

3. A Robust Approach to Multi-Stage Stochastic Linear Optimization

We now present the proposed data-driven approach, based on robust optimization, for solving multi-stage stochastic linear optimization. First, we construct an uncertainty set $\mathcal{U}_N^j \subseteq \Xi$ around each sample path, consisting of realizations $\boldsymbol{\zeta} \equiv (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T)$ which are slight perturbations of $\hat{\boldsymbol{\xi}}^j \equiv (\hat{\boldsymbol{\xi}}_1^j, \ldots, \hat{\boldsymbol{\xi}}_T^j)$. Then, we optimize for decision rules by averaging over the worst-case costs from each uncertainty set, and require that the decision rule is feasible for all realizations in all of the uncertainty sets. Formally, the proposed approach is the following:

$$\begin{aligned} \underset{\mathbf{x} \in \mathcal{X}}{\text{minimize}} \quad & \frac{1}{N} \sum_{j=1}^N \sup_{\boldsymbol{\zeta} \in \mathcal{U}_N^j} \sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\zeta}) \cdot \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \\ \text{subject to} \quad & \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\zeta}) \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \le \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j. \end{aligned} \tag{2}$$

In contrast to traditional robust optimization, Problem (2) involves averaging over multiple uncertainty sets. Thus, the explicit goal here is to obtain decision rules 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 encompass 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

$$\mathcal{U}_N^j \coloneqq \left\{ \boldsymbol{\zeta} \equiv (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T) \in \Xi : \; \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}^j\| \le \epsilon_N \right\},$$

where $\epsilon_N \geq 0$ is a parameter which controls the size of the uncertainty sets. The parameter is indexed by N to allow for the size of the uncertainty sets to change as more data is obtained. 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 provides a near-optimal approximation of Problem (1) in the presence of big data (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

$$\begin{aligned} &\text{minimize} && \mathbb{E}\left[x_2(\xi_1) + 2x_3(\xi_1,\xi_2)\right] \\ &\text{subject to} && x_2(\xi_1) + x_3(\xi_1,\xi_2) \geq \xi_1 + \xi_2 \quad \text{a.s.} \\ &&& x_2(\xi_1),\; x_3(\xi_1,\xi_2) \geq 0 \qquad\quad\;\; \text{a.s.} \end{aligned} \tag{3}$$

The output of the optimization problem are decision rules, $x_2 : \mathbb{R} \to \mathbb{R}$ 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 $(\hat{\xi}_1^1, \hat{\xi}_2^1), \ldots, (\hat{\xi}_1^N, \hat{\xi}_2^N)$. 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

$$\begin{aligned} \text{minimize} \quad & \frac{1}{N} \sum_{j=1}^{N} \left(x_2(\hat{\xi}_1^j) + 2x_3(\hat{\xi}_1^j, \hat{\xi}_2^j) \right) \\ \text{subject to} \quad & x_2(\hat{\xi}_1^j) + x_3(\hat{\xi}_1^j, \hat{\xi}_2^j) \geq \hat{\xi}_1^j + \hat{\xi}_2^j \qquad \forall j \in [N] \\ & x_2(\hat{\xi}_1^j),\ x_3(\hat{\xi}_1^j, \hat{\xi}_2^j) \geq 0 \qquad \forall j \in [N]. \end{aligned}$$

Suppose that the random variable ξ_1 for preorders has a continuous distribution. In that case, it immediately follows that $\hat{\xi}_1^1 \neq \cdots \neq \hat{\xi}_1^N$ almost surely, and thus an optimal decision rule for the above optimization problem is

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

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). $\square$

In the following section, we show that the asymptotic overfitting phenomenon illustrated in the above example is eliminated by adding robustness to the historical data via Problem (2).

4. Asymptotic Optimality

In this section, we provide asymptotic bounds on the gap between the optimal cost of our robust optimization approach, Problem (2), and the optimal cost of the multi-stage stochastic linear optimization problem, Problem (1), as the number of sample paths grows to infinity. From a practical standpoint, the bounds suggest that our robust optimization problem can provide a reasonable approximation of the multi-stage stochastic linear optimization problem in the presence of big data. Moreover, we show that our bounds collapse in particular examples of multi-stage stochastic linear optimization, in which case our robust optimization approach is guaranteed to be asymptotically optimal. These results can be viewed as significant due to the generality of multi-stage stochastic optimization problems considered in this paper.

In Section 4.1, we describe the assumptions used in the subsequent convergence results. In Section 4.2, we present the main result of this paper (Theorem 1), which establishes asymptotic lower and bound bounds on our proposed data-driven approach. In Section 4.3, we interpret Theorem 1 through several examples. In Section 4.4, we present asymptotic feasibility guarantees.

4.1. Assumptions

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

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

For example, this assumption is satisfied if the stochastic process has a multivariate Gaussian distribution, and is not satisfied if the stochastic process has a multivariate exponential distribution. Importantly, Assumption 1 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 ϵ_N is chosen to be strictly positive and decreases to zero as more data is obtained at the following rate:

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

In a nutshell, Assumption 2 provides a theoretical requirement on how to choose the robustness parameter to ensure that Problem (2) will not overfit the historical data (see Example 1 from Section 3). The rate also 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 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:*

$$\sup_{\boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j} \left\|\mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1})\right\| \leq \sup_{\boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j} L(1 + \|\boldsymbol{\zeta}\|) \quad \forall t \in [T].$$

This assumption says that there always exists 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 we find can be easily verified in many practical examples. In Appendix A, we show that every example presented in this paper satisfies this assumption.

4.2. Main result

We now present the main result of this paper (Theorem 1), which establishes asymptotic lower and upper bounds on the optimal cost of Problem (2). For notational convenience, let J^* be the optimal cost of Problem (1), $\widehat{J}_N$ be the optimal cost of Problem (2), and $S \subseteq \Xi$ be the support of the underlying joint probability distribution of the stochastic process.

First, let $\underline{J}$ be defined as the maximal optimal cost of any chance-constrained variant of the multi-stage stochastic linear optimization problem:

$$\begin{aligned}
\underline{J} := \quad \lim_{\rho \downarrow 0} \quad \underset{\mathbf{x} \in \mathcal{X},\, \tilde{S} \subseteq \Xi}{\text{minimize}} \quad & \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \mathbb{I}\left\{\boldsymbol{\xi} \in \tilde{S}\right\}\right] \\
\text{subject to} \quad & \sum_{t=1}^{T} \mathbf{A}_t(\boldsymbol{\zeta}) \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \leq \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \tilde{S} \\
& \mathbb{P}\left(\boldsymbol{\xi} \in \tilde{S}\right) \geq 1 - \rho.
\end{aligned}$$

We observe that the above limit must exist, as the optimal cost of the chance-constrained optimization problem is monotone in ρ. We also observe that $\underline{J}$ is always a lower bound on J^*, since for every $\rho > 0$, adding the constraint $\mathbb{P}(\boldsymbol{\xi} \in \tilde{S}) = 1$ on the decision variable $\tilde{S}$ to the above chance-constrained optimization problem would increase its optimal cost to J^*.[1]

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:

$$\begin{aligned}
\bar{J} := \quad \lim_{\rho \downarrow 0} \quad \underset{\mathbf{x} \in \mathcal{X}}{\text{minimize}} \quad & \bar{\mathbb{E}}\left[\sum_{t=1}^{T} \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1})\right] \\
\text{subject to} \quad & \sum_{t=1}^{T} \mathbf{A}_t(\boldsymbol{\zeta}) \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \leq \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \Xi : \operatorname{dist}(\boldsymbol{\zeta}, S) \leq \rho.
\end{aligned}$$

We remark that the limit as ρ tends down to zero must exist as well, since the optimal cost of the above optimization problem with expanded support is monotone in ρ. Note also that the expectation in the objective function has been replaced with $\bar{\mathbb{E}}[\cdot]$, which we define here as the local upper semicontinuous envelope of an expectation, *i.e.*, $\bar{\mathbb{E}}[f(\boldsymbol{\xi})] := \lim_{\epsilon \to 0} \mathbb{E}[\sup_{\boldsymbol{\zeta} \in \Xi : \|\boldsymbol{\zeta} - \boldsymbol{\xi}\| \leq \epsilon} f(\boldsymbol{\zeta})]$. 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 1. *Suppose Assumptions 1, 2, and 3 hold. Then,* $\mathbb{P}^\infty$*-almost surely we have*

$$\underline{J} \leq \liminf_{N \to \infty} \widehat{J}_N \leq \limsup_{N \to \infty} \widehat{J}_N \leq \bar{J}.$$

[1] The definition does not preclude the possibility that $\underline{J}$ is equal to $-\infty$ or ∞. However, we do not expect either of those values to occur outside of pathological cases; see Section 4.3. The same remark applies to the upper bound $\bar{J}$.

Proof. See Appendix C. □

Note that Theorem 1 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, these asymptotic bounds for Problem (2) do not necessitate imposing any restrictions on the space of decision rules. To the best of our knowledge, such nonparametric convergence guarantees for a sample-path approach to multi-stage stochastic linear optimization are the first of their kind when uncertainty is correlated across time.

Our proof of Theorem 1 is based on a new uniform convergence result (Theorem 2) which establishes a general relationship for arbitrary functions between the in-sample worst-case cost and the expected out-of-sample cost over the uncertainty sets. We state this theorem below due to its independent interest.

THEOREM 2. *If Assumptions 1 and 2 hold, then there exists a $\bar{N} \in \mathbb{N}$, $\mathbb{P}^\infty$-almost surely, such that*

$$\mathbb{E}\left[f(\boldsymbol{\xi})\mathbb{I}\left\{\boldsymbol{\xi} \in \cup_{j=1}^N \mathcal{U}_N^j\right\}\right] \le \frac{1}{N}\sum_{j=1}^N \sup_{\boldsymbol{\zeta} \in \mathcal{U}_N^j} f(\boldsymbol{\zeta}) + M_N \sup_{\boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j} |f(\boldsymbol{\zeta})|$$

for all $N \ge \bar{N}$ and all measurable functions $f : \mathbb{R}^d \to \mathbb{R}$, where $M_N := N^{-\frac{1}{(d+1)(d+2)}} \log N$.

Proof. See Appendix D. □

We note that our proofs of Theorems 1 and 2 also utilize a feasibility guarantee (Theorem 3) which can be found Section 4.4.

4.3. Examples where $\bar{J} - \underline{J}$ is zero or strictly positive

Theorem 1 establishes asymptotic lower and upper bounds on the optimal cost of Problem (2). We next show that the lower and upper bounds can be equal, $\underline{J} = \bar{J}$, in which case the optimal cost of Problem (2) provably converges to the optimal cost of Problem (1). We first show that these bounds can be equal by revisiting the stochastic inventory management problem from Example 1.

PROPOSITION 1. *For Problem* (3), *$\underline{J} = J^*$. If there is an optimal $x_2^* : \mathbb{R} \to \mathbb{R}$ for Problem* (3) *which is continuous, then $\bar{J} = J^*$.*

Proof. See Appendix B.

The proof of this proposition holds for any underlying probability distribution which satisfies Assumption 1. In combination with Theorem 1, the above proposition shows that adding robustness to the historical data provably overcomes the asymptotic overfitting phenomenon discussed in Section 3.

More generally, the equality of the lower bound $\underline{J}$ and upper bound $\bar{J}$ can be established for much broader classes of problems than Example 1. In Appendix B, we provide sufficient conditions for the lower and upper bounds to be equal and demonstrate that these sufficient conditions for asymptotic optimality can be satisfied in the real-world applications studied in Sections 7 and 8.

Unless restrictions are placed on the space of multi-stage stochastic linear optimization problems, we show next that the lower and upper bounds can have a nonzero gap. In the following, we present three examples that provide intuition on the situations in which this gap may be strictly positive. Our first example presents a problem in which the lower bound $\underline{J}$ is equal to J^* but is strictly less than the upper bound $\bar{J}$.

EXAMPLE 2. Consider the single-stage stochastic problem

$$\begin{aligned} &\underset{x_1 \in \mathbb{Z}}{\text{minimize}} && x_1 \\ &\text{subject to} && x_1 \geq \xi_1 \quad \text{a.s.}, \end{aligned}$$

where the random variable ξ_1 is governed by the probability distribution $\mathbb{P}(\xi_1 > \alpha) = (1-\alpha)^k$ for fixed $k > 0$, and $\Xi = [0,2]$. We observe that the support of the random variable is $S = [0,1]$, and thus the optimal cost of the stochastic problem is $J^* = 1$. We similarly observe that the lower bound is $\underline{J} = 1$ and the upper bound, due to the integrality of the first stage decision, is $\bar{J} = 2$. If $\epsilon_N = N^{-\frac{1}{3}}$, then we prove in Appendix E that the bounds in Theorem 1 are tight under different choices of k:

Range of k	Result
$k \in (0,3)$	$\mathbb{P}^\infty\left(\underline{J} < \liminf_{N\to\infty} \widehat{J}_N = \limsup_{N\to\infty} \widehat{J}_N = \bar{J}\right) = 1$
$k = 3$	$\mathbb{P}^\infty\left(\underline{J} = \liminf_{N\to\infty} \widehat{J}_N < \limsup_{N\to\infty} \widehat{J}_N = \bar{J}\right) = 1$
$k \in (3,\infty)$	$\mathbb{P}^\infty\left(\underline{J} = \liminf_{N\to\infty} \widehat{J}_N = \limsup_{N\to\infty} \widehat{J}_N < \bar{J}\right) = 1$

This example shows that gaps can arise between the lower and upper bounds when mild changes in the support of the underlying probability distribution lead to significant changes in the optimal cost of Problem (1). Moreover, this example illustrates that each of the inequalities in Theorem 1 can hold with equality or strict inequality when the feasibility of decisions depends on random variables that have not yet been realized. □

Our second example presents a problem in which the upper bound $\bar{J}$ is equal to J^* but is strictly greater than the lower bound $\underline{J}$. This example deals with the special case in which any chance constrained version of a stochastic problem leads to a decision which is infeasible for the true stochastic problem.

Example 3. Consider the single-stage stochastic problem

$$\begin{aligned} \underset{\mathbf{x}_1 \in \mathbb{R}^2}{\text{minimize}} \quad & x_{12} \\ \text{subject to} \quad & \xi_1(1 - x_{12}) \le x_{11} \quad \text{a.s.} \\ & 0 \le x_{12} \le 1, \end{aligned}$$

where $\xi_1 \sim \text{Gaussian}(0,1)$ and $\Xi = \mathbb{R}$. The constraints are satisfied only if $x_{12} = 1$, and so the optimal cost of the stochastic problem is $J^* = 1$. Since there is no expectation in the objective and Ξ equals the true support, we also observe that $\bar{J} = 1$. However, we readily observe that there is always a feasible solution to the sample robust optimization problem (Problem (2)) where $x_{12} = 0$, and therefore $\underline{J} = \widehat{J}_N = 0$ for all $N \in \mathbb{N}$. □

Our third and final example demonstrates the necessity of the upper semicontinuous envelope $\bar{\mathbb{E}}[\cdot]$ in the definition of the upper bound.

Example 4. Consider the two-stage stochastic problem

$$\begin{aligned} \underset{x_2 : \mathbb{R} \to \mathbb{Z}}{\text{minimize}} \quad & \mathbb{E}\left[x_2(\xi_1)\right] \\ \text{subject to} \quad & x_2(\xi_1) \ge \xi_1 \quad \text{a.s.,} \end{aligned}$$

where $\theta \sim \text{Bernoulli}(0.5)$ and $\psi \sim \text{Uniform}(0,1)$ are independent random variables, $\xi_1 = \theta\psi$, and $\Xi = [0,1]$. An optimal decision rule $x_2^* : \mathbb{R} \to \mathbb{Z}$ to the stochastic problem is given by $x_2^*(\xi_1) = 0$ for all $\xi_1 \le 0$ and $x_2^*(\xi_1) = 1$ for all $\xi_1 > 0$, which implies that $J^* = \frac{1}{2}$. It follows from similar reasoning that $\underline{J} = \frac{1}{2}$. Since Ξ equals the support of the random variable, the only difference between the stochastic problem and the upper bound is that the latter optimizes over the local upper semicontinuous envelope, and we observe that $\lim_{N \to \infty} \widehat{J}_N = \bar{J} = \bar{\mathbb{E}}[x_2^*(\xi_1)] = 1$. □

In each of the above examples, we observe that the bounds in Theorem 1 are tight, in the sense that the optimal cost of Problem (2) converges either to the lower bound or the upper bound. This provides some indication that the bounds in Theorem 1 offer an accurate depiction of how Problem (2) can behave in the asymptotic regime. On the other hand, the above examples which illustrate a nonzero gap seem to require intricate construction, and future work may identify (sub-classes) of Problem (1) where the equality of the bounds can be ensured.

4.4. Feasibility guarantees

We conclude Section 4 by discussing out-of-sample feasibility guarantees for decision rules obtained from 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 leverage classic results from detection theory.

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

$$\sum_{t=1}^{T} \mathbf{A}_t(\boldsymbol{\zeta})\mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \le \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in S_N.$$

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,

$$\mathbb{P}\left(\sum_{t=1}^{T} \mathbf{A}_t(\boldsymbol{\xi})\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1}) \le \mathbf{b}(\boldsymbol{\xi})\right) \ge \mathbb{P}\left(\boldsymbol{\xi} \in S_N\right),$$

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

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 for the support S of a 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 connection between stochastic optimization and support estimation, we obtain the following guarantee on feasibility.

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

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

Proof. See Appendix F. □

Intuitively speaking, Theorem 3 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 robustness parameter ϵ_N is set to zero.

5. Approximation Techniques

In the previous section, we developed theoretical guarantees which demonstrated that Problem (2) provides a good approximation of multi-stage stochastic linear optimization when the number of sample paths is large. In this section, we demonstrate that Problem (2) can be addressed using approximation techniques from the field of robust optimization. Specifically, we show that two decision-rule approximation schemes from robust optimization, *linear decision rules* and *finite adaptability*, can be extended to obtain approximations of Problem (2). In particular, we present a novel duality argument (Theorem 4) which allows the computational cost of these techniques to scale efficiently in the number of sample paths. The computational tractability and out-of-sample performance of these approximation schemes is illustrated via numerical experiments in Sections 7 and 8.

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 Section 1.1, 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, to approximately solve 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

$$\mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) = \mathbf{x}_{t,0} + \sum_{s=1}^{t-1} \mathbf{X}_{t,s} \boldsymbol{\zeta}_s.$$

Thus, rather than optimizing over the space of all possible decision rules (functions), we instead optimize over a finite collection of decision variables which parameterize a linear decision rule. For the setting where $\mathbf{c}_t(\boldsymbol{\xi})$ and $\mathbf{A}_t(\boldsymbol{\xi})$ 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

$$\begin{aligned} &\text{minimize} && \frac{1}{N} \sum_{j=1}^{N} \sup_{\boldsymbol{\zeta} \in \mathcal{U}_N^j} \sum_{t=1}^{T} \mathbf{c}_t \cdot \left(\mathbf{x}_{t,0} + \sum_{s=1}^{t-1} \mathbf{X}_{t,s} \boldsymbol{\zeta}_s \right) \\ &\text{subject to} && \sum_{t=1}^{T} \mathbf{A}_t \left(\mathbf{x}_{t,0} + \sum_{s=1}^{t-1} \mathbf{X}_{t,s} \boldsymbol{\zeta}_s \right) \le \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \cup_{j=1}^{N} \mathcal{U}_N^j, \end{aligned} \tag{4}$$

where the decision variables are $\mathbf{x}_{t,0} \in \mathbb{R}^{n_t}$ and $\mathbf{X}_{t,s} \in \mathbb{R}^{n_t \times d_s}$ for all $1 \le s < t \le T$ and the affine function $\mathbf{b}(\boldsymbol{\zeta}) \in \mathbb{R}^m$ is shorthand for $\mathbf{b}^0 + \sum_{t=1}^{T} \mathbf{B}_t \boldsymbol{\zeta}_t$.

Much like linear decision rules in robust optimization, we observe that Problem (4), when feasible, always produces a feasible decision rule 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) to 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 linear optimization with size that scales linearly in the number of sample paths (Theorem 4). 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* $\mathcal{U}_N^j := \{\boldsymbol{\zeta} \in \mathbb{R}^d : \boldsymbol{\ell}^j \le \boldsymbol{\zeta} \le \mathbf{u}^j\}$.

For example, Assumption 4 holds if we choose the set Ξ to be $\mathbb{R}^d_+$ and use the $\|\cdot\|_\infty$ norm in the uncertainty sets from Section 3. The following illustrates the novel duality technique described above:

Theorem 4. *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

$$\begin{aligned} \sum_{t=1}^{T} \left(\sum_{s=t+1}^{T} \mathbf{X}_{s,t}^\intercal \mathbf{c}_s \right) \cdot \boldsymbol{\zeta}_t &\le v_j - \sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{x}_{t,0} && \forall \boldsymbol{\zeta} \in \mathcal{U}_N^j,\ j \in \{1, \ldots, N\}, \\ \sum_{t=1}^{T} \left(-\mathbf{B}_t + \sum_{s=t+1}^{T} \mathbf{A}_s \mathbf{X}_{t,s} \right) \boldsymbol{\zeta}_t &\le \mathbf{b}^0 - \sum_{t=1}^{T} \mathbf{A}_t \mathbf{x}_{t,0} && \forall \boldsymbol{\zeta} \in \mathcal{U}_N^j,\ j \in \{1, \ldots, N\}. \end{aligned}$$

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

$$\max_{\boldsymbol{\zeta} \in \mathcal{U}_N^j} \sum_{t=1}^{T} \mathbf{d}_t \cdot \boldsymbol{\zeta}_t \le \gamma$$

for some vector $\mathbf{d} \coloneqq (\mathbf{d}_1, \ldots, \mathbf{d}_T) \in \mathbb{R}^d$ and scalar $\gamma \in \mathbb{R}$. Moreover, it follows from strong duality for linear optimization that

$$\max_{\boldsymbol{\zeta} \in \mathcal{U}_N^j} \sum_{t=1}^T \mathbf{d}_t \cdot \boldsymbol{\zeta}_t \quad = \quad \begin{cases} \underset{\boldsymbol{\mu}_t, \boldsymbol{\lambda}_t \in \mathbb{R}_+^{d_t}}{\text{minimize}} & \sum_{t=1}^T \left(\mathbf{u}_t^j \cdot \boldsymbol{\mu}_t - \boldsymbol{\ell}_t^j \cdot \boldsymbol{\lambda}_t \right) \\ \text{subject to} & \boldsymbol{\mu}_t - \boldsymbol{\lambda}_t = \mathbf{d}_t \qquad \forall t \in [T], \end{cases}$$

where $\mathbf{u}^j \coloneqq (\mathbf{u}_1^j, \ldots, \mathbf{u}_T^j) \in \mathbb{R}^d$ and $\boldsymbol{\ell}^j \coloneqq (\boldsymbol{\ell}_1^j, \ldots, \boldsymbol{\ell}_T^j) \in \mathbb{R}^d$ are the upper and lower bounds which define the uncertainty set. We readily observe that the solutions $\boldsymbol{\mu}_t = [\mathbf{d}_t]_+$ and $\boldsymbol{\lambda}_t = [-\mathbf{d}_t]_+$ 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 $\boldsymbol{\alpha} \coloneqq (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T) \in \mathbb{R}_+^d$ and $\boldsymbol{\beta} \coloneqq (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T) \in \mathbb{R}_+^d$ which satisfy

$$\begin{aligned} \sum_{t=1}^T \left(\boldsymbol{\alpha}_t \cdot \mathbf{u}_t^j - \boldsymbol{\beta}_t \cdot \boldsymbol{\ell}_t^j + \mathbf{c}_t \cdot \mathbf{x}_{t,0} \right) \le v_j \quad & \forall j \in [N] \\ \boldsymbol{\alpha}_t - \boldsymbol{\beta}_t = \sum_{s=t+1}^T \mathbf{X}_{s,t}^\intercal \mathbf{c}_s \quad & \forall t \in [T] \end{aligned}$$

and there exists $\mathbf{M} \coloneqq (\mathbf{M}_1, \ldots, \mathbf{M}_T) \in \mathbb{R}_+^{m \times d}$ and $\boldsymbol{\Lambda} \coloneqq (\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_T) \in \mathbb{R}_+^{m \times d}$ which satisfy

$$\begin{aligned} \sum_{t=1}^T \left(\mathbf{M}_t \mathbf{u}_t^j - \boldsymbol{\Lambda}_t \boldsymbol{\ell}_t^j + \mathbf{A}_t \mathbf{x}_{t,0} \right) \le \mathbf{b}^0 \quad & \forall j \in [N] \\ \mathbf{M}_t - \boldsymbol{\Lambda}_t = -\mathbf{B}_t + \sum_{s=t+1}^T \mathbf{A}_s \mathbf{X}_{t,s} \quad & \forall t \in [T] \end{aligned}$$

Removing the epigraph decision variables, the resulting reformulation of Problem (4) is

$$\begin{aligned} \text{minimize} \quad & \frac{1}{N} \sum_{j=1}^N \sum_{t=1}^T \left(\boldsymbol{\alpha}_t \cdot \mathbf{u}_t^j - \boldsymbol{\beta}_t \cdot \boldsymbol{\ell}_t^j + \mathbf{c}_t \cdot \mathbf{x}_{t,0} \right) \\ \text{subject to} \quad & \boldsymbol{\alpha}_t - \boldsymbol{\beta}_t = \sum_{s=t+1}^T \mathbf{X}_{s,t}^\intercal \mathbf{c}_s & t \in [T] \\ & \sum_{t=1}^T \left(\mathbf{M}_t \mathbf{u}_t^j - \boldsymbol{\Lambda}_t \boldsymbol{\ell}_t^j + \mathbf{A}_t \mathbf{x}_{t,0} \right) \le \mathbf{b}^0 & j \in [N] \\ & \mathbf{M}_t - \boldsymbol{\Lambda}_t = -\mathbf{B}_t + \sum_{s=t+1}^T \mathbf{A}_s \mathbf{X}_{t,s} & t \in [T] \end{aligned}$$

where the auxiliary decision variables are $\boldsymbol{\alpha} \equiv (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_T), \boldsymbol{\beta} \equiv (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_T) \in \mathbb{R}_+^d$ and $\mathbf{M} \equiv (\mathbf{M}_1, \ldots, \mathbf{M}_T), \boldsymbol{\Lambda} \equiv (\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_T) \in \mathbb{R}_+^{m \times d}$. Thus, the reformulation technique allowed us to decrease the number of auxiliary decision variables from $O(Nmd)$ to $O(md)$. $\square$

While linear decision rules can sometimes provide a near-optimal approximation of Problem (2) (see Section 8), we do not expect this to be the case in general. Indeed, we recall from Section 4 that

Problem (2) can provide a near-optimal approximation of Problem (1), 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). 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 (see Section 5.2 for an example). Moreover, Problem (2) is also amenable to new approximation schemes which exploit its particular structure; we refer to our companion paper Bertsimas et al. (2021) 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 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 partitions the set Ξ into smaller regions, and then optimizes a separate static or linear decision rule 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 4) readily extend to this richer class of decision rules, and a practical demonstration of finite adaptability is presented 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 $P_t^k \subseteq \mathbb{R}^{d_1+\cdots+d_t}$ 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

$$\mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) = \begin{cases} \mathbf{x}_{t,0}^1 + \sum_{s=1}^{t-1} \mathbf{X}_{t,s}^1 \boldsymbol{\zeta}_s, & \text{if } (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \in P_{t-1}^1, \\ \vdots \\ \mathbf{x}_{t,0}^K + \sum_{s=1}^{t-1} \mathbf{X}_{t,s}^K \boldsymbol{\zeta}_s, & \text{if } (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \in P_{t-1}^K. \end{cases}$$

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 $\mathbf{x}_{t,0}^k$ to be integer and restrict the associated rows of each matrix $\mathbf{X}_{t,s}^k$ 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 $(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1})$, 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 $\mathbf{x}^k_t = \mathbf{x}^{k'}_t$ and $\mathbf{X}^k_{t,s} = \mathbf{X}^{k'}_{t,s}$ for every $1 \le s < t$ when the regions P^k and $P^{k'}$ are indistinguishable at stage t.

Proposition 2 **(Proposition 4, Bertsimas and Dunning (2016))**. *If there exists $\boldsymbol{\zeta} \equiv (\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_T) \in P^k$ and $\boldsymbol{\zeta}' \equiv (\boldsymbol{\zeta}'_1,\ldots,\boldsymbol{\zeta}'_T) \in P^{k'}$ such that $(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) = (\boldsymbol{\zeta}'_1,\ldots,\boldsymbol{\zeta}'_{t-1})$, and $\boldsymbol{\zeta} \in \operatorname{int}(P^{k'})$ or $\boldsymbol{\zeta}' \in \operatorname{int}(P^k)$ hold, then we must enforce the constraints that $\mathbf{x}^k_{t,0} = \mathbf{x}^{k'}_{t,0}$ and $\mathbf{X}^k_{t,s} = \mathbf{X}^{k'}_{t,s}$ for all $1 \le s < t$ 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 Ξ, and let the intersections between regions of the partition and uncertainty sets be denoted by $\mathcal{K}^j := \{k \in [K] : \mathcal{U}^j_N \cap P^k \neq \emptyset\}$. For the setting where $\mathbf{c}_t(\boldsymbol{\xi})$ and $\mathbf{A}_t(\boldsymbol{\xi})$ do not depend on the uncertain parameters, the resulting linear decision rule approximation of Problem (2) is given by

$$\begin{aligned}
&\text{minimize} && \frac{1}{N}\sum_{j=1}^N \max_{k \in \mathcal{K}^j} \max_{\boldsymbol{\zeta} \in \mathcal{U}^j_N \cap P^k} \sum_{t=1}^T \mathbf{c}_t \cdot \left(\mathbf{x}^k_{t,0} + \sum_{s=1}^{t-1} \mathbf{X}^k_{t,s}\boldsymbol{\zeta}_s\right) \\
&\text{subject to} && \sum_{t=1}^T \mathbf{A}_t \left(\mathbf{x}^k_{t,0} + \sum_{s=1}^{t-1} \mathbf{X}^k_{t,s}\boldsymbol{\zeta}_s\right) \le \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}^j_N \cap P^k,\ k \in [K] \\
& && \mathbf{x}^k_t = \mathbf{x}^{k'}_t,\ \mathbf{X}^k_{t,s} = \mathbf{X}^{k'}_{t,s} \qquad \forall (k,k',t) \in \mathcal{T}(P^1,\ldots,P^K),\ 1 \le s < t.
\end{aligned} \tag{5}$$

where the decision variables are $\mathbf{x}^k_{t,0} \in \mathbb{R}^{n_t}$ and $\mathbf{X}^k_{t,s} \in \mathbb{R}^{n_t \times d_s}$ for all $1 \le s < t$ and $k \in [K]$.

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) can produce a tighter approximation of Problem (2), although this comes with an increase in problem size. For heuristic algorithms for selecting the partitions, we refer the reader to Postek and Den Hertog (2016) and Bertsimas and Dunning (2016). Once the partitions are determined, we obtain a reformulation of Problem (5) by employing the same duality techniques as in Section 5.1. To this end, 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 $\mathcal{U}^j_N \cap P^k := \{\boldsymbol{\zeta} \in \mathbb{R}^d : \boldsymbol{\ell}^{jk} \le \boldsymbol{\zeta} \le \mathbf{u}^{jk}\}$ 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 show that Problem (5) can 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.

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

$$
\begin{aligned}
&\text{minimize} && \frac{1}{N}\sum_{j=1}^{N} v_j \\
&\text{subject to} && \sum_{t=1}^{T}\left(\mathbf{u}_t^{jk}\cdot\boldsymbol{\alpha}_t^k-\boldsymbol{\ell}_t^{jk}\cdot\boldsymbol{\beta}_t^k+\mathbf{c}_t\cdot\mathbf{x}_{t,0}^k\right)\le v_j && j\in[N],\ k\in\mathcal{K}_j \\
& && \boldsymbol{\alpha}_t^k-\boldsymbol{\beta}_t^k=\sum_{s=t+1}^{T}\left(\mathbf{X}_{s,t}^k\right)^{\intercal}\mathbf{c}_s && t\in[T],\ k\in[K] \\
& && \sum_{t=1}^{T}\left(\mathbf{M}_t^k\mathbf{u}_t^{jk}-\boldsymbol{\Lambda}_t^k\boldsymbol{\ell}_t^{jk}+\mathbf{A}_t\mathbf{x}_{t,0}^k\right)\le\mathbf{b}^0 && j\in[N],\ k\in\mathcal{K}_j \\
& && \mathbf{M}_t^k-\boldsymbol{\Lambda}_t^k=-\mathbf{B}_t+\sum_{s=t+1}^{T}\mathbf{A}_s\mathbf{X}_{t,s}^k && t\in[T],\ k\in[K] \\
& && \mathbf{x}_t^k=\mathbf{x}_t^{k'},\ \mathbf{X}_{t,s}^k=\mathbf{X}_{t,s}^{k'} && (k,k',t)\in\mathcal{T}(P^1,\ldots,P^K),\ 1\le s<t,
\end{aligned}
$$

where the auxiliary decision variables are $\mathbf{v}\in\mathbb{R}^N$ *as well as* $\boldsymbol{\alpha}^k:=(\boldsymbol{\alpha}_1^k,\ldots,\boldsymbol{\alpha}_T^k),\boldsymbol{\beta}^k:=(\boldsymbol{\beta}_1^k,\ldots,\boldsymbol{\beta}_T^k)\in\mathbb{R}_+^d$ *and* $\mathbf{M}^k:=(\mathbf{M}_1^k,\ldots,\mathbf{M}_T^k),\boldsymbol{\Lambda}^k:=(\boldsymbol{\Lambda}_1^k,\ldots,\boldsymbol{\Lambda}_T^k)\in\mathbb{R}_+^{m\times d}$ *for each* $k\in[K]$*. Note that* $\mathbf{b}(\boldsymbol{\zeta}):=\mathbf{b}^0+\sum_{t=1}^{T}\mathbf{B}_t\boldsymbol{\zeta}_t\in\mathbb{R}^m$.

Proof. The proof follows from similar duality techniques as Theorem 4 and is thus omitted.

This result suggests that Problem (2) with finite adaptability is scalable, in the sense that the size of the resulting reformulation for a given partition $P^1,\ldots,P^K$ scales lightly in the number of sample paths N. Assuming that the partition's regions and uncertainty sets are hyperrectangles, we remark that $\boldsymbol{\ell}^{jk},\mathbf{u}^{jk}$, and $\mathcal{T}(P^1,\ldots,P^K)$ can be obtained 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 distributionally robust optimization, and discuss implications of our results to the latter.

Our exposition in this section focuses on the following formulation of multi-stage distributionally robust linear optimization:

$$\begin{aligned} &\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} && \sup_{\mathbb{Q}\in\mathcal{A}_N} \mathbb{E}_{\mathbb{Q}}\left[\sum_{t=1}^{T} \mathbf{c}_t(\boldsymbol{\xi})\cdot\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})\right] \\ &\text{subject to} && \sum_{t=1}^{T} \mathbf{A}_t(\boldsymbol{\xi})\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1}) \le \mathbf{b}(\boldsymbol{\xi}) \quad \mathbb{Q}\text{-a.s., } \forall \mathbb{Q}\in\mathcal{A}_N. \end{aligned} \tag{6}$$

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. (2019) 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 ∞-Wasserstein distance is defined as

$$\mathsf{d}_\infty(\mathbb{Q},\mathbb{Q}') := \inf\left\{ \Pi\text{-}\underset{\Xi\times\Xi}{\operatorname{ess\,sup}} \|\boldsymbol{\xi}-\boldsymbol{\xi}'\| : \begin{array}{c} \Pi \text{ is a joint distribution of } \boldsymbol{\xi} \text{ and } \boldsymbol{\xi}' \\ \text{with marginals } \mathbb{Q} \text{ and } \mathbb{Q}', \text{ respectively} \end{array} \right\},$$

where the essential supremum of the joint distribution is given by

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

From an intuitive standpoint, we note, in the case of $d=1$, that the ∞-Wasserstein distance between two bounded probability distributions can be interpreted as the maximum distance between the quantile functions of the two distributions; see Ramdas et al. (2017). For any $p\in[1,\infty)$, the p-Wasserstein distance between two probability distributions is defined as

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

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

$$\mathcal{A}_N = \left\{ \mathbb{Q}\in\mathcal{P}(\Xi) : \mathsf{d}_p\left(\mathbb{Q},\widehat{\mathbb{P}}_N\right) \le \epsilon_N \right\},$$

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

As discussed at the end of Section 1.1, there are relatively few previous convergence guarantees for distributionally robust optimization with the ∞-Wasserstein ambiguity set, even for single-stage problems. Indeed, when the underlying distribution is unbounded, the ∞-Wasserstein ambiguity set will never contain the true distribution, even as N tends to infinity, since the distance $\mathsf{d}_\infty(\mathbb{P}, \widehat{\mathbb{P}}_N)$ from the true to the empirical distribution will always be infinite. Thus, except under stronger assumptions than Assumption 1, the techniques used by Mohajerin 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 extend to the ∞-Wasserstein ambiguity set. Nonetheless, distributionally robust optimization with the ∞-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 proposition shows that Problem (2), under the particular construction of uncertainty sets from Section 3, can also be interpreted as Problem (6) with the ∞-Wasserstein ambiguity set.

Proposition 3. *Problem* (2) *with uncertainty sets of the form*

$$\mathcal{U}_N^j := \left\{ \boldsymbol{\zeta} \equiv (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T) \in \Xi : \ \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}^j\| \leq \epsilon_N \right\}$$

is equivalent to ∞*-WDRO.*

Proof. See Appendix G. $\square$

Therefore, as a byproduct of Theorem 1 from Section 4, we have obtained general convergence guarantees for distributionally robust optimization using the ∞-Wasserstein ambiguity set under mild 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 proposition shows that the constraints induced by such an approach are overly conservative in general.

Proposition 4. *If* $p \in [1, \infty)$ *and* $\epsilon_N > 0$*, then a decision rule is feasible for* p*-WDRO only if*

$$\sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\zeta}) \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \leq \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \Xi.$$

Proof. See Appendix H. $\square$

As discussed in Section 2, the set Ξ 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 with $p \in [1,\infty)$ 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 Mohajerin 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 lead to suboptimal decisions, even for problems where uncertainty does not impact feasibility, if the true support is not known exactly. For example, consider the problem

$$\begin{aligned} \underset{x_2:\mathbb{R}\to\mathbb{R},\, x_3:\mathbb{R}^2\to\mathbb{R}}{\text{minimize}} \quad & \mathbb{E}\left[x_2(\xi_1) + 2x_3(\xi_1,\xi_2)\right] \\ \text{subject to} \quad & x_2(\xi_1) + x_3(\xi_1,\xi_2) \ge \xi_1 + \xi_2 \quad \text{a.s.} \\ & x_2(\xi_1) + x_3(\xi_1,\xi_2) \ge \xi_1 - \xi_2 \quad \text{a.s.} \end{aligned}$$

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 (ξ_1,ξ_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

$$\begin{aligned} \underset{x_{2,0},\, x_{2,1},\, x_{3,0},\, x_{3,1},\, x_{3,2} \in \mathbb{R}}{\text{minimize}} \quad & \sup_{\mathbb{Q} \in \mathcal{A}_N} \mathbb{E}_{\mathbb{Q}}\left[(x_{2,0} + x_{2,1}\xi_1) + 2(x_{3,0} + x_{3,1}\xi_1 + x_{3,2}\xi_2)\right] \\ \text{subject to} \quad & (x_{2,0} + x_{2,1}\zeta_1) + (x_{3,0} + x_{3,1}\zeta_1 + x_{3,2}\zeta_2) \ge \zeta_1 + \zeta_2 \qquad \forall \boldsymbol{\zeta} \in \mathbb{R}^2 \\ & (x_{2,0} + x_{2,1}\zeta_1) + (x_{3,0} + x_{3,1}\zeta_1 + x_{3,2}\zeta_2) \ge \zeta_1 - \zeta_2 \qquad \forall \boldsymbol{\zeta} \in \mathbb{R}^2. \end{aligned}$$

It follows from identical reasoning as Bertsimas et al. (2019, 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. A similar example is found in Section 8. □

7. Application to a Stochastic Inventory Replenishment Problem

In our first set of numerical experiments, we consider a stochastic inventory replenishment problem in a network with a single warehouse and multiple retailers. Our setting is motivated by the real-world case study of Avrahami et al. (2014), who study a supply chain controlled by a major Israeli publisher that sells a weekly magazine. The goal of the publisher is to find production and replenishment policies for delivering magazines to retailers which minimize their expected weekly operating costs. We perform numerical experiments to assess the practical value of our robust optimization approach in this application in comparison to alternative data-driven approaches from the literature.

7.1. Problem description

The stochastic inventory replenishment problem faced by the publisher can be modeled as a three-stage stochastic linear optimization problem with *mixed-integer* decision variables and a *multi-dimensional* stochastic process. Let the number of retailers in the supply chain network be denoted by $R \in \mathbb{N}$. The time horizon of the decision problem faced by the publisher corresponds to one calendar week, beginning on Sunday and ending on Saturday, in which procurement and replenishment decisions are made at the start and middle of the week. The dynamics of the decision problem are summarized below:

- (Sunday) At the beginning of the week, the publisher decides the production quantity and the destinations of the weekly magazine. These decisions are captured by the decision variables $Q_{11}, \ldots, Q_{1R} \geq 0$, which represent the number of produced magazines delivered directly to each of the retailers at the start of the week, and the decision variable $Q_{10} \geq 0$, which represents the number of produced magazines sent to a common warehouse. Magazines are produced at a cost of c per unit, and $\sum_{r=0}^{R} Q_{1r}$ is the total number of magazines that are produced for the entire week.

- (Sunday - Tuesday) After the publisher has made the initial production decisions, each retailer $r \in \{1, \ldots, R\}$ uses their inventory of Q_{1r} magazines to satisfy the random customer demand ξ_{1r} from the first half of the week. There is no backlogging for demand that exceeds the available inventory, and the inventory at each retailer at the end of the first half of the week is thus equal to $\max\{0, Q_{1r} - \xi_{1r}\}$. The publisher incurs a cost of b per unit of unmet customer demand at each of the retailers, and the realized demands from the first half of the week $\boldsymbol{\xi}_1 := (\xi_{11}, \ldots, \xi_{1R})$ are observed by the publisher on Tuesday night.

- (Wednesday) At the midweek point, the publisher can replenish the inventories of the retailers using the magazines from the warehouse. These replenishment quantities to each of the retailers are captured by second-stage decision rules, $Q_{21}(\boldsymbol{\xi}_1), \ldots, Q_{2R}(\boldsymbol{\xi}_1) \geq 0$. The replenishment quantities are constrained by the inventory at the warehouse, $\sum_{r=1}^{R} Q_{2r}(\boldsymbol{\xi}_1) \leq Q_{10}$, and the publisher pays a fixed shipping cost $f \geq 0$ for each retailer to which it sends a nonzero replenishment quantity.

- (Wednesday - Saturday) After the publisher has made the replenishment decisions, each retailer $r \in \{1, \ldots, R\}$ uses their inventory of $\max\{0, Q_{1r} - \xi_{1r}\} + Q_{2r}(\boldsymbol{\xi}_1)$ magazines to satisfy the random customer demand ξ_{2r} from the second half of the week. The random demands in the second half of the week at each of the retailers are represented by $\boldsymbol{\xi}_2 := (\xi_{21}, \ldots, \xi_{2R})$.

- (Saturday) At the end of the week, the publisher incurs a holding cost of h per unit and a backlogging cost of b per unit at each of the retailers, and a holding cost of h per unit is also applied to any remaining units at the warehouse.

It follows from the above dynamics that the stochastic inventory replenishment problem faced by the publisher corresponds to a three-stage stochastic nonlinear optimization problem of the form

$$\begin{aligned}
\underset{\mathbf{Q}\geq\mathbf{0},\mathbf{z}\in\{0,1\}^R,\mathbf{I}}{\text{minimize}} \quad & \mathbb{E}\Bigg[c\left(Q_{10}+\sum_{r=1}^R Q_{1r}\right)+h\left(Q_{10}-\sum_{r=1}^R Q_{2r}(\boldsymbol{\xi}_1)\right)+b\left(\sum_{r=1}^R \max\{0,\xi_{1r}-Q_{1r}\}\right)\\
& \quad +f\left(\sum_{r=1}^R z_r(\boldsymbol{\xi}_1)\right)+b\left(\sum_{r=1}^R \max\{0,-I_{3r}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)\}\right)+h\left(\sum_{r=1}^R \max\{0,I_{3r}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)\}\right)\Bigg]\\
\text{subject to} \quad & \sum_{r=1}^R Q_{2r}(\boldsymbol{\xi}_1)\leq Q_{10} && \text{a.s.}\\
& I_{3r}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)=\max\{0,Q_{1r}-\xi_{1r}\}+Q_{2r}(\boldsymbol{\xi}_1)-\xi_{2r} && \forall r\in[R],\ \text{a.s.}\\
& z_r(\boldsymbol{\xi}_1)\mathcal{M}\geq Q_{2r}(\boldsymbol{\xi}_1) && \forall r\in[R],\ \text{a.s.},
\end{aligned}$$

where the decision rules $I_{31}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2),\ldots,I_{3R}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)$ represent the net inventory at each retailer at the end of the week, the decision rules $z_1(\boldsymbol{\xi}_1),\ldots,z_R(\boldsymbol{\xi}_1)\in\{0,1\}$ represent whether a fixed cost should be applied at each retailer, and $\mathcal{M}$ is a sufficiently large big-M constant. Following similar reasoning as Avrahami et al. (2014, Appendix A), we show in Appendix J.1 that the above optimization problem can be equivalently reformulated as the following three-stage stochastic linear optimization problem:

$$\begin{aligned}
\underset{\mathbf{Q}\geq\mathbf{0},\mathbf{z}\in\{0,1\}^R,\mathbf{v}}{\text{minimize}} \quad & \mathbb{E}\left[c\left(Q_{10}+\sum_{r=1}^R Q_{1r}\right)+hQ_{10}+\sum_{r=1}^R v_r(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)+f\sum_{r=1}^R z_r(\boldsymbol{\xi}_1)\right]\\
\text{subject to} \quad & \sum_{r=1}^R Q_{2r}(\boldsymbol{\xi}_1)\leq Q_{10} && \text{a.s.}\\
& v_r(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)\geq b(\xi_{2r}+\xi_{1r}-Q_{2r}(\boldsymbol{\xi}_1)-Q_{1r})-hQ_{2r}(\boldsymbol{\xi}_1) && \forall r\in[R],\ \text{a.s.}\\
& v_r(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)\geq h(Q_{1r}-\xi_{1r}-\xi_{2r}) && \forall r\in[R],\ \text{a.s.}\\
& v_r(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)\geq b(\xi_{1r}-Q_{1r})-h\xi_{2r} && \forall r\in[R],\ \text{a.s.}\\
& z_r(\boldsymbol{\xi}_1)\mathcal{M}\geq Q_{2r}(\boldsymbol{\xi}_1) && \forall r\in[R],\ \text{a.s.}
\end{aligned}\tag{7}$$

7.2. Experiments

We compare several approaches for finding decision rules for Problem (7) when the only information on the joint probability distribution of the stochastic process comes from historical data. Specifically, following the discussion of Avrahami et al. (2014), we assume that the publisher has collected historical data consisting of the demands for magazines in past weeks, $\hat{\boldsymbol{\xi}}^1 := (\hat{\boldsymbol{\xi}}_1^1,\hat{\boldsymbol{\xi}}_2^1),\ldots,\hat{\boldsymbol{\xi}}^N := (\hat{\boldsymbol{\xi}}_1^N,\hat{\boldsymbol{\xi}}_2^N)$, which are independent and identically distributed sample paths of the underlying stochastic process $\boldsymbol{\xi} := (\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)$. We also assume that the random demands are known to be nonnegative almost surely, $\Xi := \mathbb{R}_+^{2R}$. We compare the following data-driven approaches for finding decision rules for Problem (7):

- *SAA - Independence*: Given that the only information on the underlying joint probability distribution comes from sample paths, Avrahami et al. (2014) study a method for obtaining approximate solutions to Problem (7) which assumes that the demands in the first half of the week $\boldsymbol{\xi}_1$ and the second half of the week $\boldsymbol{\xi}_2$ are independent random vectors. The assumption of stage-wise independence is a common simplifying assumption in the stochastic programming literature which allows for a scenario tree to be constructed directly from historical sample paths. Specifically, under this assumption of stage-wise independence, this approach finds first-stage decisions for Problem (7) by solving the following mixed-integer linear optimization problem constructed from the scenario tree:

$$
\begin{aligned}
\underset{\mathbf{Q}\geq \mathbf{0},\mathbf{z}\in\{0,1\}^{N\times R},\mathbf{v}}{\text{minimize}} \quad & c\left(Q_{10}+\sum_{r=1}^{R} Q_{1r}\right)+hQ_{10}+\frac{1}{N}\sum_{j=1}^{N}\sum_{r=1}^{R}\left(fz_r^j+\frac{1}{N}\sum_{k=1}^{N} v_r^{jk}\right) \\
\text{subject to} \quad & \sum_{r=1}^{R} Q_{2r}^j \leq Q_{10} && j\in[N] \\
& v_r^{jk} \geq b\left(\hat{\xi}_{2r}^k+\hat{\xi}_{1r}^j-Q_{2r}^j-Q_{1r}\right)-hQ_{2r}^j && j,k\in[N],\ r\in[R] \\
& v_r^{jk} \geq h\left(Q_{1r}-\hat{\xi}_{1r}^j-\hat{\xi}_{2r}^k\right) && j,k\in[N],\ r\in[R] \\
& v_r^{jk} \geq b\left(\hat{\xi}_{1r}^j-Q_{1r}\right)-h\hat{\xi}_{2r}^k && j,k\in[N],\ r\in[R] \\
& z_r^j\mathcal{M}\geq Q_{2r}^j && j\in[N],\ r\in[R].
\end{aligned}
\tag{8}
$$

We denote the optimal first-stage decisions to the above linear optimization problem by $\hat{Q}_{10}^{\text{SAA}},\hat{Q}_{11}^{\text{SAA}}\ldots,\hat{Q}_{1R}^{\text{SAA}}\geq 0$. Given these first-stage decisions and a realization of the demand in the first half of the week, $\bar{\boldsymbol{\xi}}_1=(\bar{\xi}_{11}\ldots,\bar{\xi}_{1R})$, the approach obtains second-stage decisions by solving the following mixed-integer linear optimization problem:

$$
\begin{aligned}
\underset{\mathbf{Q}_2\geq \mathbf{0},\mathbf{z}\in\{0,1\}^{R},\mathbf{v}}{\text{minimize}} \quad & \sum_{r=1}^{R}\left(fz_r+\frac{1}{N}\sum_{k=1}^{N} v_r^k\right) \\
\text{subject to} \quad & \sum_{r=1}^{R} Q_{2r}\leq \hat{Q}_{10}^{\text{SAA}} \\
& v_r^k\geq b\left(\hat{\xi}_{2r}^k+\bar{\xi}_{1r}-Q_{2r}-\hat{Q}_{1r}^{\text{SAA}}\right)-hQ_{2r} && k\in[N],\ r\in[R] \\
& v_r^k\geq h\left(\hat{Q}_{1r}^{\text{SAA}}-\bar{\xi}_{1r}-\hat{\xi}_{2r}^k\right) && k\in[N],\ r\in[R] \\
& v_r^k\geq b\left(\bar{\xi}_{1r}-\hat{Q}_{1r}^{\text{SAA}}\right)-h\hat{\xi}_{2r}^k && k\in[N],\ r\in[R] \\
& z_r\mathcal{M}\geq Q_{2r} && r\in[R].
\end{aligned}
$$

We denote the optimal decisions to the above optimization problem by $\hat{Q}_{21}^{\text{SAA}}(\bar{\boldsymbol{\xi}}_1),\ldots,\hat{Q}_{2R}^{\text{SAA}}(\bar{\boldsymbol{\xi}}_1)\geq 0$ and $\hat{z}_1^{\text{SAA}}(\bar{\boldsymbol{\xi}}_1),\ldots,\hat{z}_R^{\text{SAA}}(\bar{\boldsymbol{\xi}}_1)\in\{0,1\}$.

- *AR Linear*: This approach obtains an approximation of Problem (7) using a parametric 'estimate-then-optimize' technique from the literature. Specifically, the approach consists of two steps: in the 'estimate' step, the historical data $\hat{\boldsymbol{\xi}}^1,\ldots,\hat{\boldsymbol{\xi}}^N$ is fit to a parametric family of joint probability distributions; in the 'optimize' step, a scenario tree is constructed by sampling from the estimated joint probability distribution, and decision rules are obtained by solving the corresponding approximation of the multi-stage stochastic linear optimization problem.

In the 'estimate' step, the scenario tree is constructed by first generating $\tilde{N}$ realizations for the demand in the first half of the week and then, for each such realization $\tilde{\boldsymbol{\xi}}_1^j$, generating a conditional set of demand realizations for the second half of the week $\{\tilde{\boldsymbol{\xi}}_2^{jk}\}_{k \in [\tilde{N}]}$. This generation procedure creates a scenario tree with $\tilde{N}^2$ leafs. Given this scenario tree, decisions are obtained by solving Problem (8) where $\hat{\xi}_{2r}^k$ is replaced by $\tilde{\xi}_{2r}^{jk}$ in each of the constraints. Following the stochastic programming literature (see, *e.g.*, Löhndorf and Shapiro (2019)), we fit the historical data to a general linear model of the form $\boldsymbol{\xi}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{\xi}_2 \sim \mathcal{N}\left(\mathbf{B}\begin{bmatrix}1 \\ \boldsymbol{\xi}_1\end{bmatrix}, \boldsymbol{\Sigma}_2\right)$, and we use the standard maximum likelihood estimators for the parameters $\boldsymbol{\mu}_1$, $\mathbf{B}$, $\boldsymbol{\Sigma}_1$, and $\boldsymbol{\Sigma}_2$ (Finn 1974, Section 4.4).

- *SRO-FA*: Our final method obtains decision rules for Problem (7) using the proposed robust optimization approach. Specifically, given historical sample paths of the form $\hat{\boldsymbol{\xi}}^1 := (\hat{\boldsymbol{\xi}}_1^1, \hat{\boldsymbol{\xi}}_2^1), \ldots, \hat{\boldsymbol{\xi}}^N := (\hat{\boldsymbol{\xi}}_1^N, \hat{\boldsymbol{\xi}}_2^N)$, we first construct an instance of Problem (2) in which the uncertainty sets from Section 3 are defined with the ℓ_∞-norm. We then approximately solve the robust optimization problem to obtain decision rules for Problem (7). To obtain a tractable approximation of the robust optimization problem, we implement the finite adaptability technique described in Section 5.2.

In our implementation of the finite adaptability, we construct a partition of $\Xi = \mathbb{R}_+^{2R}$ in which the historical demands in the first half of the week, $\hat{\boldsymbol{\xi}}_1^1, \ldots, \hat{\boldsymbol{\xi}}_1^N$, each lie in their own hyperrectangular region. In other words, we construct a partition of Ξ consisting of regions of the form $P^1 := [\boldsymbol{\ell}^1, \mathbf{u}^1] \times \mathbb{R}_+^R, \ldots, P^N := [\boldsymbol{\ell}^N, \mathbf{u}^N] \times \mathbb{R}_+^R$ which satisfy the property that $\hat{\boldsymbol{\xi}}^j \in P^j$ for each sample path $j \in [N]$. This partition is motivated by the desire to obtain a tight approximation of the robust optimization problem; indeed, we observe that the number and granularity of regions will increase with the number of sample paths. Details on the partitioning heuristic used in this section can be found in Appendix J.2.

Given a partition of the form described above, we approximate the robust optimization problem, Problem (2), by restricting the space of second-stage decision rules to those of the form

$$Q_{2r}(\boldsymbol{\zeta}_1) = \begin{cases} Q_{2r}^1, & \text{if } \boldsymbol{\ell}^1 \le \boldsymbol{\zeta}_1 \le \mathbf{u}^1, \\ \vdots \\ Q_{2r}^N, & \text{if } \boldsymbol{\ell}^N \le \boldsymbol{\zeta}_1 \le \mathbf{u}^N; \end{cases} \qquad z_r(\boldsymbol{\zeta}_1) = \begin{cases} z_r^1, & \text{if } \boldsymbol{\ell}^1 \le \boldsymbol{\zeta}_1 \le \mathbf{u}^1, \\ \vdots \\ z_r^N, & \text{if } \boldsymbol{\ell}^N \le \boldsymbol{\zeta}_1 \le \mathbf{u}^N. \end{cases}$$

For notational convenience, let $\mathcal{K}_j := \{k \in [N] : \mathcal{U}_N^j \cap P^k \neq \emptyset\}$ denote the indices of regions $P^1, \ldots, P^N$ that intersect the uncertainty set $\mathcal{U}_N^j$, and let us define the quantities $\underline{\zeta}_{tr}^{jk} := \min_{\boldsymbol{\zeta} \in \mathcal{U}_N^j \cap P^k} \zeta_{tr}$ and $\bar{\zeta}_{tr}^{jk} := \max_{\boldsymbol{\zeta} \in \mathcal{U}_N^j \cap P^k} \zeta_{tr}$ for each period $t \in \{1, 2\}$, retailer $r \in \{1, \ldots, R\}$, sample path $j \in \{1, \dots, N\}$, and region $k \in \mathcal{K}_j$. With this notation, we show in Appendix J.3 that the resulting approximation of Problem (2) using finite adaptability is obtained by solving the following mixed-integer linear optimization problem:

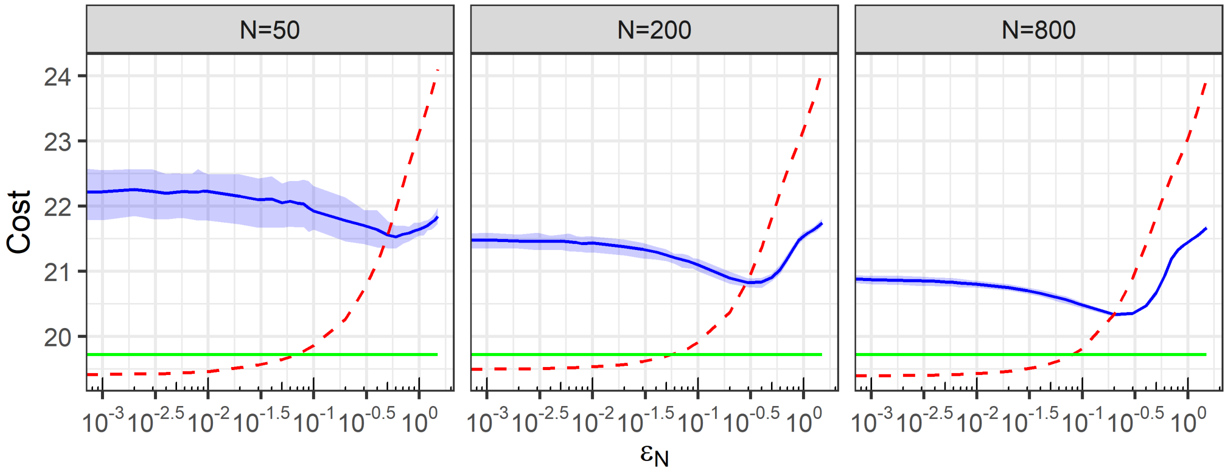
$$
\begin{aligned}
\underset{\substack{\mathbf{v}^j,\, \mathbf{Q}_1 \geq \mathbf{0}, \\ \mathbf{Q}_2^k \geq 0, \mathbf{z}^k \in \{0,1\}^R \\ \mathbf{u}^{j,k}, \mathbf{v}^j \in \mathbb{R}^R}}{\text{minimize}} \quad & c\left(Q_{10} + \sum_{r=1}^R Q_{1r}\right) + hQ_{10} + \frac{1}{N}\sum_{j=1}^N \sum_{r=1}^R v_r^j \\
\text{subject to} \quad & \sum_{r=1}^R Q_{2r}^k \leq Q_{10} && \forall k \in [K] \\
& v_r^j \geq \sum_{r=1}^R \left(u_r^{j,k} + f z_r^k\right) && \forall r \in [R],\, j \in [N],\, k \in \mathcal{K}_j \\
& u_r^{j,k} \geq b(\bar{\zeta}_{2r}^{jk} + \bar{\zeta}_{1r}^{jk} - Q_{2r}^k - Q_{1r}) - hQ_{2r}^k && \forall r \in [R],\, j \in [N],\, k \in \mathcal{K}_j \\
& u_r^{j,k} \geq h(Q_{1r} - \underline{\zeta}_{1r}^{jk} - \underline{\zeta}_{2r}^{jk}) && \forall r \in [R],\, j \in [N],\, k \in \mathcal{K}_j \\
& u_r^{j,k} \geq b\left(\bar{\zeta}_{1r}^{jk} - Q_{1r}\right) - h\underline{\zeta}_{2r}^{jk} && \forall r \in [R],\, j \in [N],\, k \in \mathcal{K}_j \\
& z_r^k \mathcal{M} \geq Q_{2r}^k && \forall r \in [R],\, k \in [K].
\end{aligned}
\tag{9}
$$

Solving the above optimization problem yields the first-stage decisions $\hat{Q}_{10}^{\text{SRO}}, \hat{Q}_{11}^{\text{SRO}} \dots, \hat{Q}_{1R}^{\text{SRO}} \geq 0$ and decision rules $\hat{Q}_{21}^{\text{SRO}}(\boldsymbol{\xi}_1), \dots, \hat{Q}_{2R}^{\text{SRO}}(\boldsymbol{\xi}_1) \geq 0$ and $\hat{z}_1^{\text{SRO}}(\boldsymbol{\xi}_1), \dots, \hat{z}_R^{\text{SRO}}(\boldsymbol{\xi}_1) \in \{0,1\}$.

To compare the above data-driven approaches, we perform a variety of numerical experiments on different numbers of retailers, $R \in \{2,3,4\}$, with cost parameters $c = 0.25$, $h = 0.05$ and $b = 0.5$, and in cases with no fixed cost $f = 0$ as well as when the fixed cost is $f = 0.1$. In the joint probability distribution of the stochastic process, the demand in the first half of the week for each retailer, ξ_{1r}, is independent of the demands in the first half of the week for the other retailers and is generated from a truncated normal distribution with mean 6 and standard deviation 2.5 that is bounded below by zero. The demands in the second half of the week are independent across retailers and follow truncated normal distributions with means $2(6 - \xi_{1r})^2$ and standard deviations 2.5 that are bounded below by zero. This joint probability distribution is chosen because of its relative simplicity and, particularly in the setting with multiple retailers, the difficulty of correctly identifying its structure using only limited historical data. The number of retailers in our experiments is motivated by the organizational structure of the Israeli publisher, in which sales representatives are assigned to managing the inventory of up to five retailers (Avrahami et al. 2014, p. 452). In Appendix J.4, we provide numerical evidence that the fixed cost of $f = 0.1$ leads to replenishment decision rules that are meaningfully different than those obtained in experiments with no fixed cost.

We compare the aforementioned data-driven methods on training sets of sizes $N \in \{50, 100, 200, 400, 800\}$ when there is no fixed cost ($f = 0$) and $N \in \{50, 71, 100, 141, 200, 282\}$ when there are fixed costs ($f = 0.1$). To obtain statistically meaningful results, for each size N, we generate

Figure 1 Three-stage inventory replenishment problem:
Impact of robustness parameter on SRO-FA for $\mathrm{R}=3$, no fixed cost ($f=0$)



Note. The solid black lines are the average out-of-sample costs of decision rules produced by SRO-FA, and the shaded regions are the 20th and 80th percentiles over the 50 training datasets. The dotted red lines are the average in-sample costs for SRO-FA. The dashed green line is the benchmark for Problem (7). Results are shown for $N \in \{50, 200, 800\}$.

50 training datasets of size N and apply each of the above data-driven approaches to each training dataset. We evaluate the decision rules from each approach on a common testing dataset of 10,000 sample paths. For "SRO-FA", we select the robustness parameter using five-fold cross-validation. For "AR Linear", we generate scenario trees of size $\tilde{N}=50$ for all sizes of training sets N. We found that larger values of $\tilde{N}$ resulted in longer computation times but did not have any discernible impact on the performance of the "AR Linear" approach.

For comparison purposes, we also implement a benchmark method that has perfect knowledge of the true joint probability distribution. In the benchmark, we obtain an estimate of the optimal cost of Problem (7) by constructing a scenario tree from the true joint probability distribution. We construct the scenario tree using the same procedure as described previously, with a size of $\tilde{N}=800$ for experiments without fixed cost and $\tilde{N}=200$ for experiments with fixed cost. We repeat this process over 50 replications and report the average of the resulting optimal costs.

7.3. Results

In Figure 1, we show the impact of the robustness parameter on the in-sample and out-of-sample cost of the decision rules obtained by our robust optimization approach. The results demonstrate that a *strictly positive* choice of the robustness parameter is essential in order to obtain the best out-of-sample cost for each N. This is due to the fact that Problem (2) has been approximated with a rich space of decision rules, and in particular, a space of decision rules that becomes increasingly flexible as more historical sample paths are obtained. In contrast, when the robustness parameter is set to zero, Figure 1 shows that the in-sample cost of Problem (9) does not appear to converge to the optimal cost of the stochastic problem; these findings are consistent with the discussion of Example 1 in Section 3 and show that approximating Problem (2) with a rich restricted space of decision rules will asymptotically overfit the historical data when the robustness parameter is set to zero. Additional numerical results on the relationship between the robustness parameter and the performance of our robust optimization approach can be found in Appendix J.4.

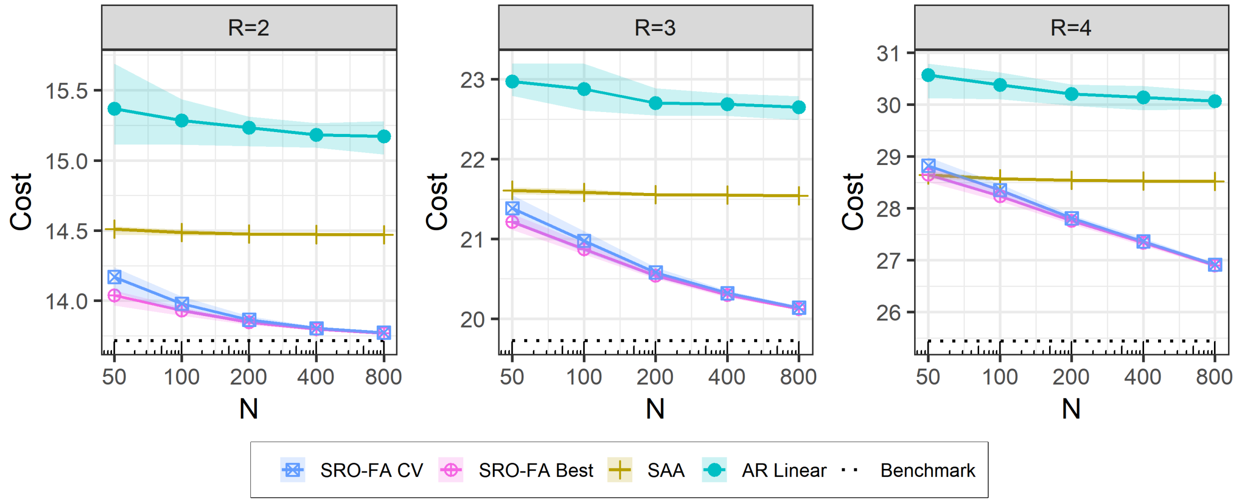
More broadly, we believe that Figure 1 highlights a practical strength of our robust optimization approach. Indeed, the approximation of Problem (2) using piecewise static decision rules did not require nor utilize any information about the structure of optimal decision rules for the underlying stochastic problem. At the same time, despite searching over a rich space of decision rules, Problem (2) with an appropriate choice of the robustness parameter yields decision rules which do not overfit the historical data. This shows that Problem (2) provides an opportunity to find high-quality decision rules for multi-stage stochastic linear optimization problems, even when (*i*) the only information on the underlying distribution comes from limited data, and (*ii*) the structure of optimal decision rules for the stochastic problem is complex or unknown.

In Figures 2 and 3, we compare the average out-of-sample costs of the decision rules produced by the various data-driven methods in experiments with no fixed cost ($f = 0$) and with fixed cost ($f = 0.1$). We first observe that the gap between "SRO-FA CV" (where the robustness parameter is chosen through five-fold cross validation) and the unrealistic "SRO-FA Best" (where the robustness parameter is chosen *post hoc* to obtain decision rules with the best average out-of-sample performance) is relatively small across sizes of training sets. This suggests that cross-validation can indeed be practically effective in choosing the robustness parameter in the robust optimization approach. Moreover, the results of the experiments show that the proposed robust optimization approach ("SRO-FA CV") can find decision rules which significantly outperform those obtained by widely-used alternative data-driven approaches ("SAA Independence" and "AR Linear"). In particular, we note that the problem sizes in our experiments are realistic, both in terms of the number of retailers R and sizes of training sets N. The results of the experiments thus provide numerical evidence that the robust optimization approach proposed in this paper can be valuable in practice, particularly in challenging applications with mixed-integer decisions and multi-dimensional stochastic processes.

8. Application to a Multi-Stage Stochastic Inventory Management Problem

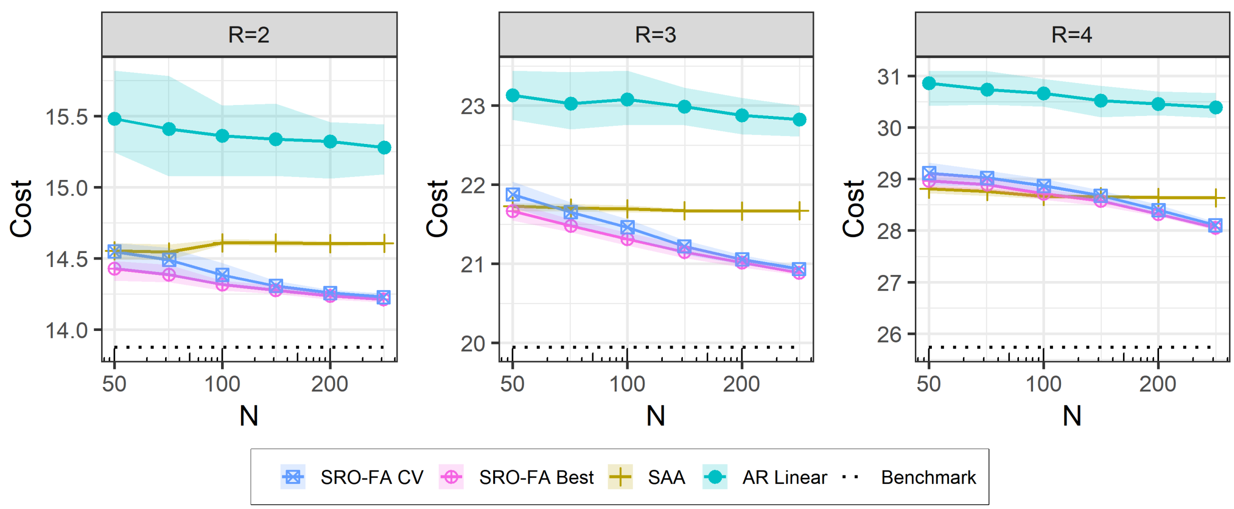
In our second set of experiments, we consider a classic and widely-studied stochastic inventory management problem for a single product with an unknown autoregressive demand. Our motivation in this

Figure 2 Three-stage inventory replenishment problem: Performance of data-driven approaches, no fixed cost ($f = 0$)



Note. The solid lines are the average out-of-sample costs of decision rules produced by the various data-driven approaches to Problem (7), and the shaded regions are the 20th and 80th percentiles over the 50 training datasets. The robustness parameter in "SRO-FA CV" is chosen using five-fold cross validation, and the robustness parameter in "SRO-FA Best" is chosen optimally with respect to the testing dataset.

Figure 3 Three-stage inventory replenishment problem: Performance of data-driven approaches, fixed cost ($f = 0.1$)



Note. The solid lines are the average out-of-sample costs of decision rules produced by the various data-driven approaches to Problem (7), and the shaded regions are the 20th and 80th percentiles over the 50 training datasets. The robustness parameter in "SRO-FA CV" is chosen using five-fold cross validation, and the robustness parameter in "SRO-FA Best" is chosen optimally with respect to the testing dataset.

set of experiments is to explore the practical value of Problem (2) in applications where the number of time periods is comparatively large (*e.g.*, $T = 10$) and the number of historical sample paths is relatively small (*e.g.*, $10 \le N \le 100$). In this context, we compare our proposed robust optimization approach with linear decision rules to a variety of alternative data-driven approaches. In contrast to the previous section, where we compared to data-driven approaches that are practically tractable in problems with mixed-integer decisions and short time horizons, this section compares to approaches that are practically tractable in problems with continuous decisions and long time horizons.

8.1. Problem description

We consider an inventory management problem of a single product over a finite planning horizon. At the beginning of each time period $t \in [T]$, we start with $I_t \in \mathbb{R}$ units of product in inventory. We then select a production quantity of $x_t \in [0, \bar{x}_t]$ with zero lead time at a cost of c_t per unit. The product demand $\xi_t \ge 0$ is then revealed, the inventory is updated to $I_{t+1} = I_t + x_t - \xi_t$, and we incur a holding cost of $h_t \max\{I_{t+1}, 0\}$ and a backorder cost of $b_t \max\{-I_{t+1}, 0\}$. We begin with zero units of inventory in the first period. Our goal is to dynamically select the production quantities to minimize the expected total cost over the planning horizon, captured by

$$
\begin{aligned}
\underset{\mathbf{x}, \mathbf{I}, \mathbf{y}}{\text{minimize}} \quad & \mathbb{E}\left[\sum_{t=1}^{T} \left(c_t x_t(\xi_1, \ldots, \xi_{t-1}) + y_{t+1}(\xi_1, \ldots, \xi_t)\right)\right] \\
\text{subject to} \quad & I_{t+1}(\xi_1, \ldots, \xi_t) = I_t(\xi_1, \ldots, \xi_{t-1}) + x_t(\xi_1, \ldots, \xi_{t-1}) - \xi_t \quad \text{a.s., } \forall t \in [T] \\
& y_{t+1}(\xi_1, \ldots, \xi_t) \ge h_t I_{t+1}(\xi_1, \ldots, \xi_t) \quad \text{a.s., } \forall t \in [T] \\
& y_{t+1}(\xi_1, \ldots, \xi_t) \ge -b_t I_{t+1}(\xi_1, \ldots, \xi_t) \quad \text{a.s., } \forall t \in [T] \\
& 0 \le x_t(\xi_1, \ldots, \xi_{t-1}) \le \bar{x}_t \quad \text{a.s., } \forall t \in [T].
\end{aligned}
\tag{10}
$$

We consider the setting where the joint probability distribution of the stochastic process $(\xi_1, \ldots, \xi_T) \in \mathbb{R}^T$ is unknown. Our only information on the distribution comes from historical data consisting of demand realizations for past products $(\hat{\xi}_1^1, \ldots, \hat{\xi}_T^1), \ldots, (\hat{\xi}_1^N, \ldots, \hat{\xi}_T^N)$, which are independent and identically distributed sample paths of the underlying stochastic process, and knowledge that the stochastic process will be contained in $\Xi = \mathbb{R}_+^T$ almost surely.

8.2. Experiments

We perform computational experiments on the following data-driven approaches for obtaining decision rules for Problem (10):

- *SRO-LDR*: This is the proposed data-driven approach for multi-stage stochastic linear optimization (Problem (2)), where the uncertainty sets are constructed as described in Section 3 with the ℓ_∞-norm. The approach is approximated using linear decision rules (see Section 5.1) and solved using the reformulation developed in Theorem 4. We choose the robustness parameter for each training dataset using five-fold cross validation, where the range of possible values considered in the cross-validation procedure is $\epsilon_N \in \{b \cdot 10^a : a \in \{-2,-1,0,1\}, b \in \{1,\ldots,9\}\}$.

- *SAA-LDR*: This is the same approach as SRO-LDR, except the robustness parameter is set to zero.
- *Approx PCM*: This is a data-driven extension of the approach developed in Bertsimas et al. (2019). In this approach, decision rules are obtained by solving a multi-stage distributionally robust optimization problem (Problem (6)) in which $\mathcal{A}_N$ is the set of joint probability distributions with the same mean and covariance as those estimated from the historical data. This distributionally robust optimization problem is solved approximately by restricting to lifted linear decision rules, as described in Bertsimas et al. (2019, Section 3).
- *DDP* and *RDDP*: This is the robust data-driven dynamic programming approach proposed by Hanasusanto and Kuhn (2013). The approach estimates cost-to-go functions by applying kernel regression to the historical sample paths. Decisions are obtained from optimizing over the cost-to-go functions, which are evaluated approximately using the algorithm described in Hanasusanto and Kuhn (2013, Section 4). Since the algorithm 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 via the lifted linear decision rules obtained by Approx PCM. The approach also requires a robustness parameter γ, which we choose to be either $\gamma = 0$ (DDP) or $\gamma = 10$ (RDDP).
- *WDRO-LDR*: Described in Section 6, this approach obtains decision rules by solving a multi-stage distributionally robust optimization problem (Problem (6)) in which $\mathcal{A}_N$ is chosen to be the 1-Wasserstein ambiguity set with the ℓ_1-norm. Similarly as SRO-LDR, the distributionally robust optimization problem is approximated using linear decision rules, which is solved using a duality-based reformulation provided in Appendix I. The robustness parameter is chosen using the same procedure as SRO-LDR.

We perform computational simulations using the same parameters and data generation as See and Sim (2010). Specifically, the demand is 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 $\bar{\varsigma} = 40$ when $T = 5$, and $\mu = 200$ and $\bar{\varsigma} = 20$ when $T = 10$. The capacities and costs are $\bar{x}_t = 260$, $c_t = 0.1$, $h_t = 0.02$ for all $t \in [T]$, $b_t = 0.2$ for all $t \in [T-1]$, and $b_T = 2$.

To compare the above data-driven approaches, we take the following steps. For various choices of N, we generate 100 training datasets of size N and obtain decision rules by applying the above data-driven approaches to each training dataset. The out-of-sample costs of the obtained decision rules are approximated using a common testing dataset of 10000 sample paths.[2] Specifically, for each sample path $(\xi_1^i, \ldots, \xi_T^i) \in \mathbb{R}^T$ in the testing dataset, we calculate production quantities $(x_1^{\mathcal{A},i,\ell}, \ldots, x_T^{\mathcal{A},i,\ell}) \in \mathbb{R}^T$ by applying the decision rule obtained from approach $\mathcal{A}$ on the ℓ-th training dataset. The out-of-sample cost of the decision rule is then approximated as

$$\frac{1}{10000} \sum_{i=1}^{10000} \sum_{t=1}^{T} \left(c_t x_t^{\mathcal{A},i,\ell} + \max \left\{ h_t I_{t+1}^{\mathcal{A},i,\ell}, -b_t I_{t+1}^{\mathcal{A},i,\ell} \right\} \right),$$

where the inventory levels $(I_1^{\mathcal{A},i,\ell}, \ldots, I_T^{\mathcal{A},i,\ell})$ are computed from the production quantities $(x_1^{\mathcal{A},i,\ell}, \ldots, x_T^{\mathcal{A},i,\ell}) \in \mathbb{R}^T$ and the test sample path $(\xi_1^i, \ldots, \xi_T^i) \in \mathbb{R}^T$. All sample paths in the training and testing datasets are drawn independently from the true joint probability distribution.

As discussed earlier, Problem (2) is not guaranteed to find decision rules which are feasible for all realizations in Ξ. Therefore, the linear decision rules obtained by SRO-LDR and SAA-LDR, when applied to sample paths in the testing dataset, may result in production quantities which exceed $\bar{x}_1, \ldots, \bar{x}_T$ or are negative. Thus, before computing the out-of-sample costs, we first project each production quantity $x_t^{\mathcal{A},i,\ell}$ onto the interval $[0, \bar{x}_t]$ to ensure it is feasible. We discuss the impact of this projection procedure at the end of the results section.

8.3. Results

In Table 1 and Figure 4, we report the out-of-sample costs and computation times of the various approaches. SRO-LDR produces an out-of-sample cost which outperforms the other approaches, most notably when the size of the training dataset is small, and requires less than one second of computation time. We note that the out-of-sample cost of SRO-LDR roughly converges to the dynamic programming (DP) estimate of the optimal cost of Problem (10), which suggests that linear decision rules provide a good approximation of the optimal production decision rules for this particular stochastic problem. The relationship between the robustness parameter and the in-sample and out-of-sample cost of SRO-LDR is shown in Figure 5.

We briefly reflect on some notable differences between SRO-LDR and the other approaches. First, the results demonstrate that a strictly positive choice of the robustness parameter is not necessary to avoid asymptotic overfitting when Problem (2) is approximated with a fixed, finite-dimensional space of decision rules; indeed, Table 1 and Figure 5 show that SAA-LDR can provide an out-of-sample cost which is similar to that of SRO-LDR for moderate to large training datasets. However, SRO-LDR produces an out-of-sample cost which significantly outperforms SAA-LDR when N is small ($N \in \{10, 25\}$). More generally, this shows that there exist regimes in which a positive choice of the

[2] For DDP and RDDP, we only evaluated on the first 1000 sample paths in the testing dataset, due to the computational cost of optimizing over the cost-to-go functions for each testing sample path.

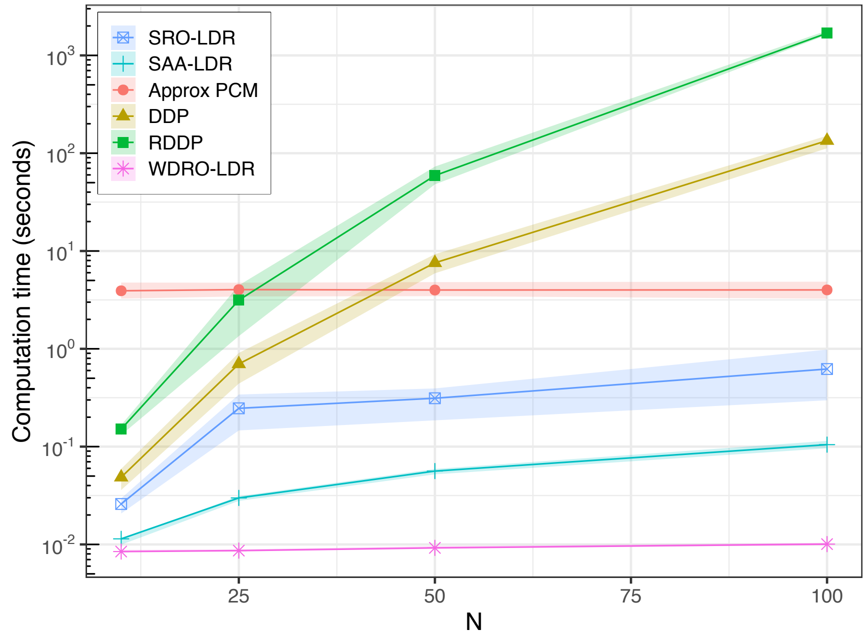
Table 1 Multi-stage stochastic inventory management: Average out-of-sample cost.

			Size of training dataset (N)				
T	α	Approach	10	25	50	100	DP
5	0	SRO-LDR	**111.8(7.2)**	**109.5(1.3)**	**108.5(0.7)**	108.0(0.3)	108
		SAA-LDR	127.3(12.6)	111.7(3.1)	108.7(1.0)	**107.9(0.3)**	
		Approx PCM	118.5(2.2)	117.4(1.0)	117.1(0.7)	117.0(0.5)	
		DDP	2262.7(363.3)	1189.6(854.2)	525.4(510.8)	205.1(201.5)	
		RDDP	2255.5(393.2)	1175.8(856.2)	515.6(506.0)	202.4(195.3)	
		WDRO-LDR	2400.3(0.0)	2400.3(0.0)	2400.3(0.0)	2400.3(0.0)	
	0.25	SRO-LDR	**113.0(4.2)**	**110.0(1.8)**	108.7(0.8)	108.0(0.3)	107
		SAA-LDR	127.6(13.0)	111.7(3.1)	**108.6(1.0)**	**108.0(0.2)**	
		Approx PCM	126.8(3.8)	125.3(1.5)	124.8(0.9)	124.6(0.8)	
		DDP	2251.5(488.8)	1393.7(897.8)	679.3(656.0)	236.9(240.5)	
		RDDP	2222.0(556.5)	1386.2(900.7)	670.1(654.9)	236.2(237.2)	
		WDRO-LDR	2400.7(0.0)	2400.7(0.0)	2400.7(0.0)	2400.7(0.0)	
	0.5	SRO-LDR	**115.5(5.3)**	**112.0(4.0)**	110.8(2.7)	111.7(2.6)	108
		SAA-LDR	129.5(13.3)	113.1(6.8)	**110.7(2.9)**	**111.6(2.6)**	
		Approx PCM	136.0(4.8)	134.0(1.9)	133.4(1.2)	133.2(1.0)	
		DDP	2263.8(480.1)	1563.7(917.6)	777.5(787.7)	364.2(488.1)	
		RDDP	2253.8(515.8)	1532.6(940.9)	716.8(758.9)	334.6(477.1)	
		WDRO-LDR	2401.2(0.0)	2401.2(0.0)	2401.2(0.0)	2401.2(0.0)	
10	0	SRO-LDR	**208.9(1.0)**	**207.5(0.6)**	**206.8(0.5)**	**206.2(0.2)**	206
		SAA-LDR	293.9(70.1)	212.6(2.1)	207.8(1.1)	206.3(0.4)	
		Approx PCM	215.3(2.1)	214.5(0.6)	214.1(0.6)	214.1(0.4)	
		DDP	5211.4(1131.1)	2827.9(1757.5)	1335.6(1206.4)	497.5(550.2)	
		RDDP	5210.1(1133.4)	2820.3(1758.7)	1327.6(1206.0)	500.1(552.7)	
		WDRO-LDR	5800.3(0.0)	5800.3(0.0)	5800.3(0.0)	5800.3(0.0)	
	0.25	SRO-LDR	**210.3(2.9)**	**207.8(1.1)**	**206.9(0.5)**	**206.3(0.2)**	206
		SAA-LDR	295.1(70.4)	212.7(2.1)	207.8(1.1)	206.3(0.4)	
		Approx PCM	228.7(4.5)	226.2(1.8)	225.7(1.1)	225.5(0.9)	
		DDP	5215.6(1350.1)	3214.7(1984.9)	1598.0(1566.5)	440.2(417.1)	
		RDDP	5202.0(1368.7)	3185.3(1977.1)	1593.6(1566.6)	437.0(418.0)	
		WDRO-LDR	5800.2(0.0)	5800.5(0.0)	5800.5(0.0)	5800.5(0.0)	
	0.5	SRO-LDR	**211.1(3.9)**	**207.9(1.0)**	**206.9(0.6)**	**206.3(0.2)**	206
		SAA-LDR	297.8(70.7)	213.0(2.3)	207.9(1.1)	206.4(0.4)	
		Approx PCM	245.3(7.0)	242.1(2.7)	241.6(2.0)	240.9(1.6)	
		DDP	5374.8(1052.7)	3676.4(2159.1)	1960.9(1878.3)	644.1(914.6)	
		RDDP	5313.0(1173.0)	3630.9(2161.4)	1949.2(1863.9)	644.0(913.3)	
		WDRO-LDR	5800.7(0.0)	5800.7(0.0)	5800.3(0.0)	5800.7(0.0)	

Mean (standard deviation) for the out-of-sample cost of decision rules obtained by various data-driven approaches for Problem (10). The robustness parameters in SRO-LDR and WDRO-LDR are chosen using cross validation. The column DP presents the dynamic programming approximations of the optimal cost of Problem (10) from See and Sim (2010, Tables EC.1 and EC.2), which have an accuracy of ±1%.

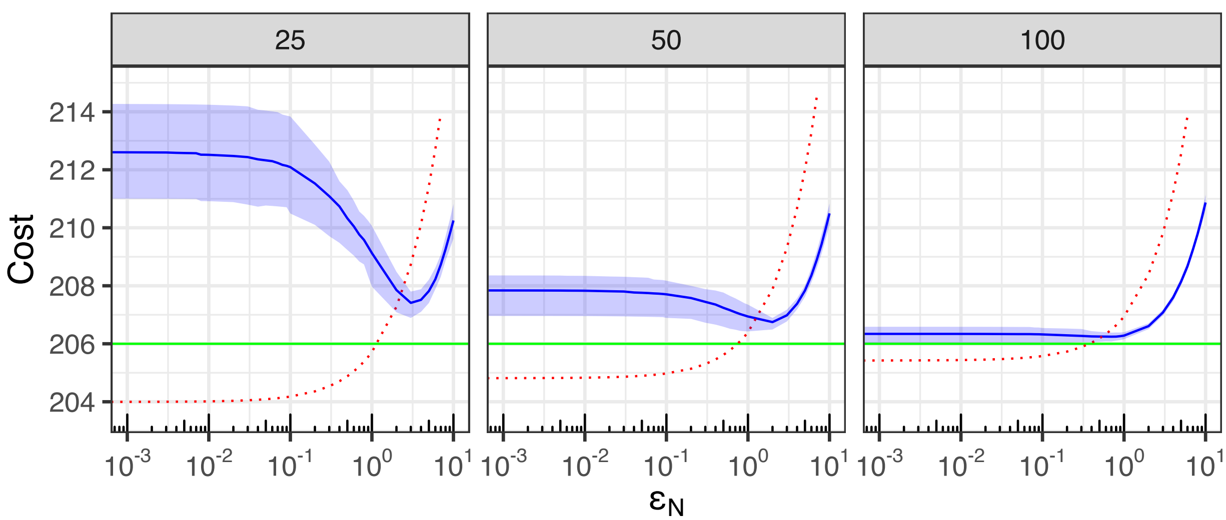
robustness parameter can still provide significant value even when Problem (2) is approximated using linear decision rules. Second, we note that WDRO-LDR consistently produces decision rules with large average out-of-sample cost; this is due to the fact that this approach requires the linear decision rules to satisfy $0 \le x_{t,0} + \sum_{s=1}^{t-1} x_{t,s}\zeta_s \le \bar{x}_t$ for all $(\zeta_1, \ldots, \zeta_T) \in \mathbb{R}_+^T$, which reduces to a static decision rule for the production quantity in each stage. Finally, we remark that the average out-of-sample cost of DDP

Figure 4 **Multi-stage stochastic inventory management: Computation times for $T = 10$, $\alpha = 0.25$.**



Note. Computation times for data-driven approaches to the multi-stage stochastic inventory management problem. Results are shown for $T = 10$ and $\alpha = 0.25$, and similar computation times were observed for other choices of α. The graph shows the mean value of the computation times over 100 training datasets for each value of N.

Figure 5 **Multi-stage stochastic inventory management:**
Impact of robustness parameter on SRO-LDR for $T = 10$, $\alpha = 0.25$.



Note. The solid black line is the average out-of-sample cost of decision rules produced by SRO-LDR, and the shaded black region is the 20th and 80th percentiles over the 100 training datasets. The dotted red line is the average in-sample cost of SRO-LDR, and the solid green line is a dynamic programming approximation of the optimal cost of Problem (10) from See and Sim (2010, Table EC.2).

and RDDP improved significantly with the size of the training dataset, but produced high variability across training datasets and required long computation time.

We recall that Table 1 reports the out-of-sample costs of SRO-LDR and SAA-LDR after their production quantities are projected onto the feasible region (see Section 8.2). In Appendix K, we discuss the impact of this projection procedure on the out-of-sample cost. Specifically, we show across the above experiments that (i) SRO-LDR produces feasible production quantities for more than 93% of the sample paths in the testing dataset, and (ii) the average ℓ_1-distance between the production quantities $(x_1^{\mathcal{A},i,\ell}, \ldots, x_T^{\mathcal{A},i,\ell})$ and the feasible region $[0, \bar{x}_1] \times \cdots \times [0, \bar{x}_T]$ is less than 2 units. This shows that SRO-LDR consistently produces feasible or nearly-feasible decisions, and thus the out-of-sample costs of SRO-LDR are unlikely to be an artifact of this projection procedure.

9. Conclusion

In this work, we presented a new data-driven approach, based on robust optimization, for solving multi-stage stochastic linear optimization problems where uncertainty is correlated across time. We showed that the proposed approach is asymptotically optimal, providing assurance that the approach offers a near-optimal approximation of the underlying stochastic problem in the presence of big data. At the same time, the optimization problem resulting from the proposed approach can be addressed by approximation algorithms and reformulation techniques which have underpinned the success of multi-stage robust optimization. The practical value of the proposed approach was illustrated by computational examples inspired by real-world applications, demonstrating that the proposed data-driven approach can produce high-quality decisions in reasonable computation times. Through these contributions, this work provides a step towards helping organizations across domains leverage historical data to make better operational decisions in dynamic environments.

References

Assaf Avrahami, Yale T Herer, and Retsef Levi. Matching supply and demand: Delayed two-phase distribution at Yedioth Group - Models, algorithms, and information technology. *Interfaces*, 44(5):445–460, 2014.

Amparo Baíllo, Antonio Cuevas, and Ana Justel. Set estimation and nonparametric detection. *Canadian Journal of Statistics*, 28(4):765–782, 2000.

Güzin Bayraksan and David K Love. Data-driven stochastic programming using phi-divergences. In *The Operations Research Revolution*, chapter 1, pages 1–19. INFORMS, 2015.

Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, 1999.

Aharon Ben-Tal, Alexander Goryashko, Elana Guslitzer, and Arkadi Nemirovski. Adjustable robust solutions of uncertain linear programs. *Mathematical Programming*, 99(2):351–376, 2004.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization.* Princeton University Press, 2009.

Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2): 341–357, 2013.

Dimitris Bertsimas and Constantine Caramanis. Finite adaptability in multistage linear optimization. *IEEE Transactions on Automatic Control*, 55(12):2751–2766, 2010.

Dimitris Bertsimas and Iain Dunning. Multistage robust mixed-integer optimization with adaptive partitions. *Operations Research*, 64(4):980–998, 2016.

Dimitris Bertsimas and Vineet Goyal. On the power of robust solutions in two-stage stochastic and adaptive optimization problems. *Mathematics of Operations Research*, 35(2):284–305, 2010.

Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011a.

Dimitris Bertsimas, Vineet Goyal, and Xu Andy Sun. A geometric characterization of the power of finite adaptability in multistage stochastic and adaptive optimization. *Mathematics of Operations Research*, 36(1):24–54, 2011b.

Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Robust sample average approximation. *Mathematical Programming*, 171(1):217–282, 2018.

Dimitris Bertsimas, Melvyn Sim, and Meilin Zhang. Adaptive distributionally robust optimization. *Management Science*, 65(2):604–618, 2019.

Dimitris Bertsimas, Shimrit Shtern, and Bradley Sturt. Two-stage sample robust optimization. *Operations Research (forthcoming)*, 2021.

John R Birge and Francois Louveaux. *Introduction to Stochastic Programming.* Springer Science & Business Media, 2011.

Leo Breiman. *Probability.* SIAM, 1992.

Xin Chen, Melvyn Sim, and Peng Sun. A robust optimization perspective on stochastic programming. *Operations Research*, 55(6):1058–1071, 2007.

George B Dantzig. Linear programming under uncertainty. *Management Science*, 1(3-4):197–206, 1955.

Erick Delage and Dan A Iancu. Robust multistage decision making. In *The Operations Research Revolution*, pages 20–46. INFORMS, 2015.

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

Luc Devroye and Gary L Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal on Applied Mathematics*, 38(3):480–488, 1980.

E. Erdoğan and Garud Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1):37–61, Jun 2006.

E. Erdoğan and Garud Iyengar. On two-stage convex chance constrained problems. *Mathematical Methods of Operations Research*, 65(1):115–140, 2007.

K Bruce Erickson. The strong law of large numbers when the mean is undefined. *Transactions of the American Mathematical Society*, 185:371–381, 1973.

Jeremy D Finn. *A general model for multivariate analysis*. Holt, Rinehart & Winston, 1974.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050v2*, 2017.

Stanley J Garstka and Roger J-B Wets. On decision rules in stochastic programming. *Mathematical Programming*, 7(1):117–143, 1974.

Angelos Georghiou, Daniel Kuhn, and Wolfram Wiesemann. The decision rule approach to optimization under uncertainty: methodology and applications. *Computational Management Science*, pages 1–32, 2018.

Angelos Georghiou, Angelos Tsoukalas, and Wolfram Wiesemann. Robust dual dynamic programming. *Operations Research*, 67(3):813–830, 2019.

Clark R Givens and Rae Michael Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

Grani A Hanasusanto and Daniel Kuhn. Robust data-driven dynamic programming. In *Advances in Neural Information Processing Systems*, pages 827–835, 2013.

Grani A Hanasusanto and Daniel Kuhn. Conic programming reformulations of two-stage distributionally robust linear programs over Wasserstein balls. *Operations Research*, 66(3):849–869, 2018.

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

Aleksandr Petrovich Korostelev and Alexandre B Tsybakov. *Minimax Theory of Image Reconstruction*, volume 82. Springer-Verlag, New York, 1993.

Pavlo Krokhmal and Stanislav Uryasev. A sample-path approach to optimal position liquidation. *Annals of Operations Research*, 152(1):193–225, 2007.

Nils Löhndorf and Alexander Shapiro. Modeling time-dependent randomness in stochastic dual dynamic programming. *European Journal of Operational Research*, 273(2):650–661, 2019.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1):115–166, 2018.

Georg Ch Pflug and Alois Pichler. From empirical observations to tree models for stochastic optimization: Convergence properties. *SIAM Journal on Optimization*, 26(3):1715–1740, 2016.

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

Krzysztof Postek and Dick Den Hertog. Multistage adjustable robust mixed-integer optimization via iterative splitting of the uncertainty set. *INFORMS Journal on Computing*, 28(3):553–574, 2016.

Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

Herbert E Scarf. A min-max solution of an inventory problem. In *Studies in the Mathematical Theory of Inventory and Production*, pages 201–209. Stanford University Press, 1958.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

Chuen-Teck See and Melvyn Sim. Robust approximation to multiperiod inventory management. *Operations Research*, 58(3):583–594, 2010.

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

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

Allen L Soyster. Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations Research*, 21(5):1154–1157, 1973.

Matthew Staib and Stefanie Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS Machine Learning and Computer Security Workshop*, 2017.

Bart PG Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 2020.

Jean-Louis Verger-Gaugry. Covering a ball with smaller equal balls in $\mathbb{R}^n$. *Discrete & Computational Geometry*, 33(1):143–155, 2005.

Hong Wang and RJ Tomkins. A zero-one law for large order statistics. *Canadian Journal of Statistics*, 20 (3):323–334, 1992.

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

Guanglin Xu and Samuel Burer. A copositive approach for two-stage adjustable robust optimization with uncertain right-hand sides. *Computational Optimization and Applications*, 70(1):33–59, 2018.

Huan Xu, Constantine Caramanis, and Shie Mannor. A distributional interpretation of robust optimization. *Mathematics of Operations Research*, 37(1):95–110, 2012.

Bo Zeng and Long Zhao. Solving two-stage robust optimization problems using a column-and-constraint generation method. *Operations Research Letters*, 41(5):457–461, 2013.

Jianzhe Zhen, Dick Den Hertog, and Melvyn Sim. Adjustable robust optimization via Fourier–Motzkin elimination. *Operations Research*, 66(4):1086–1100, 2018.

Appendix A: Verifying Assumption 3 in Examples

In this appendix, we show that each 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

$$\begin{aligned}
\underset{x_2:\mathbb{R}\to\mathbb{R},\, x_3:\mathbb{R}^2\to\mathbb{R}}{\text{minimize}} \quad & \frac{1}{N}\sum_{j=1}^{N} \sup_{\boldsymbol{\zeta}\in\mathcal{U}_N^j} \{x_2(\zeta_1) + 2x_3(\zeta_1,\zeta_2)\} \\
\text{subject to} \quad & x_2(\zeta_1) + x_3(\zeta_1,\zeta_2) \ge \zeta_1 + \zeta_2 && \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j \\
& x_2(\zeta_1),\, x_3(\zeta_1,\zeta_2) \ge 0 && \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j.
\end{aligned}$$

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:

$$\begin{aligned}
x_2(\zeta_1) &\le \sup_{\boldsymbol{\zeta}'\in\cup_{j=1}^N\mathcal{U}_N^j} \{\zeta_1' + \zeta_2'\} \quad \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j, \\
x_3(\zeta_1,\zeta_2) &\le \sup_{\boldsymbol{\zeta}'\in\cup_{j=1}^N\mathcal{U}_N^j} \{\zeta_1' + \zeta_2'\} \quad \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j.
\end{aligned}$$

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

$$\begin{aligned}
\underset{x_1\in\mathbb{Z}}{\text{minimize}} \quad & x_1 \\
\text{subject to} \quad & x_1 \ge \zeta_1 \quad \forall \zeta_1 \in \cup_{j=1}^N \mathcal{U}_N^j.
\end{aligned}$$

We observe that an optimal solution to this problem is $x_1 = \max_{\zeta_1\in\cup_{j=1}^N\mathcal{U}_N^j} \lceil \zeta_1 \rceil$, and thus the constraint

$$x_1 \le \max_{\zeta_1\in\cup_{j=1}^N\mathcal{U}_N^j} \zeta_1 + 1$$

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

A.3. Example 3 from Section 4.3

Consider the sample robust optimization problem

$$\begin{aligned}
\underset{\mathbf{x}_1\in\mathbb{R}^2}{\text{minimize}} \quad & x_{12} \\
\text{subject to} \quad & \zeta_1(1-x_{12}) \le x_{11} \quad \forall \zeta_1 \in \cup_{j=1}^N \mathcal{U}_N^j \\
& 0 \le x_{12} \le 1.
\end{aligned}$$

We observe that an optimal solution to this problem is given by $x_{11} = \max_{\zeta_1 \in \cup_{j=1}^N \mathcal{U}_N^j} \zeta_1$ and $x_{12} = 0$. Thus, the constraint

$$x_{11} \le \max_{\zeta_1 \in \cup_{j=1}^N \mathcal{U}_N^j} \zeta_1$$

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

A.4. Example 4 from Section 4.3

Consider the sample robust optimization problem

$$\begin{aligned} &\underset{x_2:\mathbb{R}\to\mathbb{Z}}{\text{minimize}} && \frac{1}{N}\sum_{j=1}^N \sup_{\zeta_1 \in \mathcal{U}_N^j} x_2(\zeta_1) \\ &\text{subject to} && x_2(\zeta_1) \ge \zeta_1 \quad \forall \zeta_1 \in \cup_{j=1}^N \mathcal{U}_N^j. \end{aligned}$$

Since $\Xi = [0,1]$, we observe that the constraint

$$x_2(\zeta_1) \le 1 \quad \forall \zeta_1 \in \cup_{j=1}^N \mathcal{U}_N^j$$

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

A.5. Example from Section 7

Consider the sample robust optimization problem associated with Problem (7):

$$\begin{aligned} &\underset{\mathbf{v},\,\mathbf{Q}_1\ge\mathbf{0},\mathbf{Q}_2,\,\mathbf{z}}{\text{minimize}} && \frac{1}{N}\sum_{j=1}^N \max_{\boldsymbol{\zeta}\in\mathcal{U}_N^j}\left\{ c\left(Q_{10} + \sum_{r=1}^R Q_{1r}\right) + hQ_{10} + \sum_{r=1}^R v_r(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2) + f\sum_{r=1}^R z_r(\boldsymbol{\zeta}_1)\right\} \\ &\text{subject to} && \sum_{r=1}^R Q_{2r}(\boldsymbol{\zeta}_1) \le Q_{10} && \forall \boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j \\ &&& v_r(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2) \ge b(\zeta_{2r} + \zeta_{1r} - Q_{2r}(\boldsymbol{\zeta}_1) - Q_{1r}) - hQ_{2r}(\boldsymbol{\zeta}_1) && \forall r\in[R],\ \boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j \\ &&& v_r(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2) \ge h(Q_{1r} - \zeta_{1r} - \zeta_{2r}) && \forall r\in[R],\ \boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j \\ &&& v_r(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2) \ge b\,(\zeta_{1r} - Q_{1r}) - h\zeta_{2r} && \forall r\in[R],\ \boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j \\ &&& z_r(\boldsymbol{\zeta}_1)\mathcal{M} \ge Q_{2r}(\boldsymbol{\zeta}_1) && \forall r\in[R],\ \boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j \\ &&& z_r(\boldsymbol{\zeta}_1)\in\{0,1\},\ Q_{2r}(\boldsymbol{\zeta}_1)\ge 0, && \forall r\in[R],\ \boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j. \end{aligned}$$

Since $\Xi = \mathbb{R}_+^{2r}$, we observe that the constraints

$$0 \le Q_{11},\dots,Q_{1R} \le \max_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j} (\zeta_{1r} + \zeta_{2r})$$

$$0 \le Q_{10} \le \max_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j} \sum_{r=1}^R (\zeta_{1r} + \zeta_{2r}),$$

can be added to the above problem without affecting its optimal cost. It thus follows from the constraint $\sum_{r=1}^R Q_{2r}(\boldsymbol{\zeta}_1) \le Q_{10}$ that the constraints

$$0 \le Q_{21}(\boldsymbol{\zeta}_1),\dots,Q_{2R}(\boldsymbol{\zeta}_1) \le \max_{\boldsymbol{\zeta}'\in\cup_{j=1}^N\mathcal{U}_N^j} \sum_{r=1}^R (\zeta'_{1r} + \zeta'_{2r}), \quad \forall \boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j$$

can also be added to the above problem without affecting its optimal cost. Finally, we can without loss of generality impose the constraint for each retailer $r \in [R]$ that

$$\begin{aligned} 0 \le v_r(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) = \max\{ & b(\zeta_{2r} + \zeta_{1r} - Q_{2r}(\boldsymbol{\zeta}_1) - Q_{1r}) - hQ_{2r}(\boldsymbol{\zeta}_1), \\ & h(Q_{1r} - \zeta_{1r} - \zeta_{2r}), \\ & b(\zeta_{1r} - Q_{1r}) - h\zeta_{2r}\} \qquad \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j. \end{aligned}$$

Applying the aforementioned bounds on the decision rules, we conclude that Assumption 3 holds.

A.6. Example from Section 8

Consider the sample robust optimization problem

$$\begin{aligned} \underset{\mathbf{x},\mathbf{I},\mathbf{y}}{\text{minimize}} \quad & \frac{1}{N}\sum_{j=1}^{T} \sup_{\boldsymbol{\zeta} \in \mathcal{U}_N^j} \sum_{t=1}^{T} \left(c_t x_t(\zeta_1,\ldots,\zeta_{t-1}) + y_{t+1}(\zeta_1,\ldots,\zeta_t)\right) \\ \text{subject to} \quad & I_{t+1}(\zeta_1,\ldots,\zeta_t) = I_t(\zeta_1,\ldots,\zeta_{t-1}) + x_t(\zeta_1,\ldots,\zeta_{t-1}) - \zeta_t \quad \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j,\ \forall t \in [T] \\ & y_{t+1}(\zeta_1,\ldots,\zeta_t) \ge h_t I_{t+1}(\zeta_1,\ldots,\zeta_t) \quad \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j,\ \forall t \in [T] \\ & y_{t+1}(\zeta_1,\ldots,\zeta_t) \ge -b_t I_{t+1}(\zeta_1,\ldots,\zeta_t) \quad \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j,\ \forall t \in [T] \\ & 0 \le x_t(\zeta_1,\ldots,\zeta_{t-1}) \le \bar{x}_t \quad \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j,\ \forall t \in [T], \end{aligned}$$

where $I_1 = 0$ and $\Xi = \mathbb{R}_+^T$. For any feasible decision rule to the above problem and for each stage t, we observe that the following constraint is satisfied:

$$-\sup_{\boldsymbol{\zeta}' \in \cup_{j=1}^N \mathcal{U}_N^j} \sum_{s=1}^{T} \zeta_s' \le I_{t+1}(\zeta_1,\ldots,\zeta_t) \le \sum_{s=1}^{T} \bar{x}_s \quad \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j.$$

Moreover, we can without loss of generality impose the constraint that

$$0 \le y_{t+1}(\zeta_1,\ldots,\zeta_t) = \max\left\{h_t I_{t+1}(\zeta_1,\ldots,\zeta_t), -b_t I_{t+1}(\zeta_1,\ldots,\zeta_t)\right\} \quad \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j.$$

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

Appendix B: On the Tightness of the Bounds from Theorem 1

In Section 4.2, we introduced a lower bound $\underline{J}$ and upper bound $\bar{J}$ on the optimal cost J^* of Problem (1). In Theorem 1, we showed under mild assumptions that these quantities also provide an asymptotic lower and upper bound on the optimal cost $\widehat{J}_N$ of Problem (2). In this appendix, we demonstrate the practical value of these bounds by establishing sufficient conditions for the lower and upper bounds to be equal and applying those sufficient conditions to applications of multi-stage stochastic linear optimization from Sections 3, 7, and 8.

B.1. Sufficient conditions for $\underline{J} = J^*$ and $\bar{J} = J^*$

We begin by developing our two primary results, Theorems EC.1 and EC.2, in which we establish sufficient conditions for the lower bound $\underline{J}$ and the upper bound $\bar{J}$ to be equal to the optimal cost J^* of Problem (1). In particular, we will show in the following Appendix B.2 that the sufficient conditions in these theorems can be verified in examples in which the underlying joint probability distribution and the support of the stochastic process are unknown. Consequently, the following two theorems can serve as practical tools for demonstrating that our proposed robust optimization approach, Problem (2), is asymptotically optimal for specific multi-stage stochastic linear optimization problems which arise in real-world applications.

We begin by developing our first primary result, Theorem EC.1, which establishes a sufficient condition for the lower bound $\underline{J}$ to be equal to the optimal cost J^* of Problem (1). Recall that S denotes the support of the joint probability distribution. Speaking intuitively, the following theorem shows that $\underline{J}$ is guaranteed to equal J^* if there exists an optimal decision rule to a stochastic problem over any restricted support $\tilde{S} \subseteq S$ that can be extended to a decision rule that is feasible for Problem (1) and has a well-behaved objective function. To see the utility of the following theorem, we refer the reader to the examples in Appendix B.2

Theorem EC.1 **(Sufficient condition for lower bound)**. *Let Assumption 1 hold, and suppose there exists an $L \geq 0$ such that, for all $\tilde{S} \subseteq S$, the optimal cost of the optimization problem*

$$\begin{aligned} \underset{\mathbf{x} \in \mathcal{X}}{\text{minimize}} \quad & \mathbb{E}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \mathbb{I}\left\{\boldsymbol{\xi} \in \tilde{S}\right\}\right] \\ \text{subject to} \quad & \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\zeta}) \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \leq \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \tilde{S} \end{aligned}$$

would not change if we added the constraints

$$\begin{aligned} 0 \leq \sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \leq L \left(1 + \max\left\{\|\boldsymbol{\xi}\|, \sup_{\boldsymbol{\zeta} \in \tilde{S}} \|\boldsymbol{\zeta}\|\right\}\right) \quad & \text{a.s.} \\ \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\xi}) \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \leq \mathbf{b}(\boldsymbol{\xi}) \quad & \text{a.s.} \end{aligned}$$

Then, $\underline{J} = J^*$.

Proof. We recall from the definition of the lower bound that $\underline{J} := \lim_{\rho \to 0} \underline{J}_\rho$, where

$$\underline{J}_\rho := \min_{\tilde{S} \subseteq \Xi:\, \mathbb{P}(\boldsymbol{\xi} \in \tilde{S}) \geq 1-\rho} \left\{ \begin{aligned} \underset{\mathbf{x} \in \mathcal{X}}{\text{minimize}} \quad & \mathbb{E}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \mathbb{I}\left\{\boldsymbol{\xi} \in \tilde{S}\right\}\right] \\ \text{subject to} \quad & \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\zeta}) \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \leq \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \tilde{S} \end{aligned} \right\}.$$

Therefore, it follows from the conditions of Theorem EC.1 that

$$\underline{J}_\rho = \min_{\tilde{S} \subseteq \Xi:\, \mathbb{P}(\boldsymbol{\xi} \in \tilde{S}) \geq 1-\rho} \left\{ \begin{aligned} \underset{\mathbf{x} \in \mathcal{X}}{\text{minimize}} \quad & \mathbb{E}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \mathbb{I}\left\{\boldsymbol{\xi} \in \tilde{S}\right\}\right] \\ \text{subject to} \quad & \sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \geq 0 \qquad \text{a.s.} \\ & \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\zeta}) \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \leq \mathbf{b}(\boldsymbol{\zeta}) \qquad \forall \boldsymbol{\zeta} \in \tilde{S} \end{aligned} \right\}.$$

For any feasible solution to the above optimization problem, we observe that the function $\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})$ is nonnegative almost surely. Therefore, for any arbitrary $r \geq 0$, a lower bound on $\underline{J}_\rho$ is given by

$$\underline{J}_{\rho,r} := \min_{\tilde{S}\subseteq\Xi:\,\mathbb{P}(\boldsymbol{\xi}\in\tilde{S})\geq 1-\rho} \left\{ \begin{array}{lll} \underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} & \mathbb{E}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi})\cdot\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})\mathbb{I}\left\{\boldsymbol{\xi}\in\tilde{S},\ \|\boldsymbol{\xi}\|\leq r\right\}\right] & \\ \text{subject to} & \sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi})\cdot\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1}) \geq 0 & \text{a.s.} \\ & \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\zeta})\mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \leq \mathbf{b}(\boldsymbol{\zeta}) & \forall \boldsymbol{\zeta}\in\tilde{S} \end{array} \right\} \tag{EC.1}$$

$$\geq \min_{\tilde{S}\subseteq\Xi:\,\mathbb{P}(\boldsymbol{\xi}\in\tilde{S})\geq 1-\rho} \left\{ \begin{array}{ll} \underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} & \mathbb{E}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi})\cdot\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})\mathbb{I}\left\{\boldsymbol{\xi}\in\tilde{S},\ \|\boldsymbol{\xi}\|\leq r\right\}\right] \\ \text{subject to} & \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\zeta})\mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \leq \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta}\in\tilde{S}: \|\boldsymbol{\zeta}\|\leq r \end{array} \right\}, \tag{EC.2}$$

where the inequality follows from removing constraints from the inner minimization problem in line (EC.1). Furthermore, it follows from the conditions of Theorem EC.1 that line (EC.2) is equal to

$$\min_{\tilde{S}\subseteq\Xi:\,\mathbb{P}(\boldsymbol{\xi}\in\tilde{S})\geq 1-\rho} \left\{ \begin{array}{lll} \underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} & \mathbb{E}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi})\cdot\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})\mathbb{I}\left\{\boldsymbol{\xi}\in\tilde{S},\ \|\boldsymbol{\xi}\|\leq r\right\}\right] & \\ \text{subject to} & \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\zeta})\mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \leq \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta}\in\tilde{S}: \|\boldsymbol{\zeta}\|\leq r & \\ & 0 \leq \sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi})\cdot\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1}) \leq L\left(1+\max\left\{\|\boldsymbol{\xi}\|, \sup_{\boldsymbol{\zeta}\in\tilde{S}:\|\boldsymbol{\zeta}\|\leq r}\|\boldsymbol{\zeta}\|\right\}\right) & \text{a.s.} \\ & \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\xi})\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1}) \leq \mathbf{b}(\boldsymbol{\xi}) & \text{a.s.} \end{array} \right\} \tag{EC.3}$$

$$\geq \min_{\tilde{S}\subseteq\Xi:\,\mathbb{P}(\boldsymbol{\xi}\in\tilde{S})\geq 1-\rho} \left\{ \begin{array}{lll} \underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} & \mathbb{E}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi})\cdot\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})\mathbb{I}\left\{\boldsymbol{\xi}\in\tilde{S},\ \|\boldsymbol{\xi}\|\leq r\right\}\right] & \\ \text{subject to} & 0 \leq \sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi})\cdot\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1}) \leq L\left(1+\max\left\{\|\boldsymbol{\xi}\|, r\right\}\right) & \text{a.s.} \\ & \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\xi})\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1}) \leq \mathbf{b}(\boldsymbol{\xi}) & \text{a.s.} \end{array} \right\}, \tag{EC.4}$$

where the inequality follows from removing constraints from the inner minimization problem in line (EC.3) and using the fact that $\sup_{\boldsymbol{\zeta}'\in\tilde{S}:\|\boldsymbol{\zeta}'\|\leq r}\|\boldsymbol{\zeta}'\| \leq r$.

We now use Assumption 1 to obtain a lower bound on (EC.4). Indeed, Assumption 1 says that there exists an $a > 1$ such that $b := \mathbb{E}[\exp(\|\boldsymbol{\xi}\|^a)] < \infty$. Therefore, for any feasible solutions $\tilde{S}\subseteq\Xi$ and $\mathbf{x}\in\mathcal{X}$ to the optimization problems in (EC.4),

$$\begin{aligned} &\mathbb{E}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi})\cdot\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})\mathbb{I}\left\{\boldsymbol{\xi}\notin\tilde{S} \text{ or } \|\boldsymbol{\xi}\|>r\right\}\right] \\ &\leq \mathbb{E}\left[L\left(1+\max\left\{\|\boldsymbol{\xi}\|, r\right\}\right)\mathbb{I}\left\{\boldsymbol{\xi}\notin\tilde{S} \text{ or } \|\boldsymbol{\xi}\|>r\right\}\right] \end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[L\left(1+\max\left\{\|\boldsymbol{\xi}\|, r\right\}\right) \mathbb{I}\left\{\boldsymbol{\xi} \notin \tilde{S}\right\}\right]+\mathbb{E}\left[L\left(1+\max\left\{\|\boldsymbol{\xi}\|, r\right\}\right) \mathbb{I}\left\{\|\boldsymbol{\xi}\|>r\right\}\right] \\
&\leq \sqrt{\mathbb{E}\left[L^2\left(1+\max\left\{\|\boldsymbol{\xi}\|, r\right\}\right)^2\right] \rho}+\sqrt{\mathbb{E}\left[L^2\left(1+\max\left\{\|\boldsymbol{\xi}\|, r\right\}\right)^2\right] \mathbb{P}(\|\boldsymbol{\xi}\|>r)} \\
&\leq \underbrace{\sqrt{\mathbb{E}\left[L^2\left(1+\max\left\{\|\boldsymbol{\xi}\|, r\right\}\right)^2\right] \rho}+\sqrt{\mathbb{E}\left[L^2\left(1+\max\left\{\|\boldsymbol{\xi}\|, r\right\}\right)^2\right] \frac{b}{\exp(r^a)}}}_{h(\rho, r)}.
\end{aligned}
$$

Indeed, the first inequality follows because $0 \leq \sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \leq L(1+\max\{\|\boldsymbol{\xi}\|, r\})$ almost surely, the second inequality follows from the union bound, the third inequality follows from $\mathbb{P}(\boldsymbol{\xi} \in \tilde{S}) \geq 1-\rho$ and the Cauchy-Schwartz inequality, and the fourth and final inequality follows from Markov's inequality. Therefore,

$$
\begin{aligned}
\underline{J}_{\rho, r} &\geq -h(\rho, r)+\min_{\tilde{S} \subseteq \Xi:\, \mathbb{P}(\boldsymbol{\xi} \in \tilde{S}) \geq 1-\rho}\left\{\begin{array}{lll}
\underset{\mathbf{x} \in \mathcal{X}}{\operatorname{minimize}} & \mathbb{E}\left[\displaystyle\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1})\right] & \\
\text{subject to} & 0 \leq \displaystyle\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \leq L\left(1+\max\left\{\|\boldsymbol{\xi}\|, r\right\}\right) & \text{a.s.} \\
& \displaystyle\sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\xi}) \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \leq \mathbf{b}(\boldsymbol{\xi}) & \text{a.s.}
\end{array}\right\} \\
&\geq -h(\rho, r)+\left\{\begin{array}{ll}
\underset{\mathbf{x} \in \mathcal{X}}{\operatorname{minimize}} & \mathbb{E}\left[\displaystyle\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1})\right] \\
\text{subject to} & \displaystyle\sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\xi}) \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \leq \mathbf{b}(\boldsymbol{\xi}) \quad \text{a.s.}
\end{array}\right. \\
&= -h(\rho, r)+J^*.
\end{aligned} \tag{EC.5}
$$

The first inequality follows from the lower bound on $\underline{J}_{\rho, r}$ in line (EC.4), the definition of $h(\rho, r)$, and the law of iterated expectation. The second inequality follows from removing constraints, and the final equality follows from the definition of J^*.

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

$$
\underline{J}=\lim_{\rho \downarrow 0} \underline{J}_\rho \geq \lim_{r \rightarrow \infty} \lim_{\rho \downarrow 0} \underline{J}_{\rho, r} \geq \lim_{r \rightarrow \infty} \lim_{\rho \downarrow 0}-h(\rho, r)+J^*=J^*.
$$

The first inequality follows because $\underline{J}_\rho \geq \underline{J}_{\rho, r}$ for any arbitrary $r \geq 0$ and the quantity $\lim_{\rho \downarrow 0} \underline{J}_{\rho, r}$ is monotonically increasing in r. The second inequality follows from (EC.5). The final equality follows from the definition of $h(\rho, r)$ and Assumption 1. Since the inequality $\underline{J} \leq J^*$ always holds, our proof is complete. □

We conclude the present Appendix B.1 by developing our second primary result, Theorem EC.2, which establishes a sufficient condition for the upper bound $\bar{J}$ to be equal to the optimal cost J^* of Problem (1). Speaking intuitively, the following theorem says that $\bar{J}$ is equal to J^* if there are near-optimal decision rules to Problem (1) that are feasible and result in an objective function which is both upper-semicontinuous and well-behaved on a slight extension of the support S.

Theorem EC.2 **(Sufficient condition for upper bound)**. *Let Assumption 1 hold, and suppose for all $\eta > 0$ that there exists an η-optimal decision rule $\mathbf{x}^\eta \in \mathcal{X}$ for Problem (1) and $\rho^\eta > 0$, $L^\eta \geq 0$ such that*

$$\sum_{t=1}^{T} \mathbf{A}_t(\boldsymbol{\zeta})\mathbf{x}_t^\eta(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \leq \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \Xi : \operatorname{dist}(\boldsymbol{\zeta}, S) \leq \rho^\eta; \tag{EC.6}$$

$$0 \leq \sum_{t=1}^{T} \mathbf{c}_t(\boldsymbol{\zeta}) \cdot \mathbf{x}_t^\eta(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \leq L^\eta(1 + \|\boldsymbol{\zeta}\|) \quad \forall \boldsymbol{\zeta} \in \Xi : \operatorname{dist}(\boldsymbol{\zeta}, S) \leq \rho^\eta; \tag{EC.7}$$

$$\lim_{\epsilon \to 0} \sup_{\boldsymbol{\zeta} \in \Xi : \, \|\boldsymbol{\zeta} - \boldsymbol{\xi}\| \leq \epsilon} \left\{ \sum_{t=1}^{T} \mathbf{c}_t(\boldsymbol{\zeta}) \cdot \mathbf{x}_t^\eta(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \right\} = \sum_{t=1}^{T} \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t^\eta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1}) \quad \text{a.s.} \tag{EC.8}$$

Then, $\bar{J} = J^*$.

Proof. For any arbitrary $\eta > 0$, consider an η-optimal decision rule $\mathbf{x}^\eta \in \mathcal{X}$ for Problem (1) and constants $\rho^\eta > 0$, $L^\eta \geq 0$ which satisfy properties (EC.6), (EC.7), and (EC.8). Then,

$$\begin{aligned}
\bar{J} &\leq \bar{\mathbb{E}}\left[\sum_{t=1}^{T} \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t^\eta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})\right] \\
&= \lim_{\epsilon \to 0} \mathbb{E}\left[\sup_{\boldsymbol{\zeta} \in \Xi : \, \|\boldsymbol{\zeta} - \boldsymbol{\xi}\| \leq \epsilon} \left\{\sum_{t=1}^{T} \mathbf{c}_t(\boldsymbol{\zeta}) \cdot \mathbf{x}_t^\eta(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1})\right\}\right] \\
&= \mathbb{E}\left[\lim_{\epsilon \to 0} \sup_{\boldsymbol{\zeta} \in \Xi : \, \|\boldsymbol{\zeta} - \boldsymbol{\xi}\| \leq \epsilon} \left\{\sum_{t=1}^{T} \mathbf{c}_t(\boldsymbol{\zeta}) \cdot \mathbf{x}_t^\eta(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1})\right\}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t^\eta(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})\right] \\
&\leq J^* + \eta.
\end{aligned}$$

Indeed, the first inequality follows from the definition of $\bar{J}$ and the feasiblility of $\mathbf{x}^\eta$ for the problem defining $\bar{J}$ for all sufficiently small ρ, as indicated by (EC.6). The first equality is the definition of the local upper semicontinuous envelope. The second equality follows from the dominated convergence theorem, which can be applied because of Assumption 1 and (EC.7). The third equality follows from (EC.8), and the final equality holds because $\mathbf{x}^\eta$ is an η-optimal decision rule for Problem (1). Since $\eta > 0$ was chosen arbitrarily and since the inequality $J^* \leq \bar{J}$ always holds, our proof is complete. $\square$

B.2. Applications of Theorems EC.1 and EC.2

In the previous section, we developed sufficient conditions, Theorems EC.1 and EC.2, for the proposed robust optimization approach, Problem (2), to be asymptotically optimal for multi-stage stochastic linear optimization problems. In this section, we use those sufficient conditions to show that Problem (2) is asymptotically optimal in three data-driven examples of multi-stage stochastic linear optimization based on Sections 3, 7, and 8. All together, the three examples provide evidence that the lower bound and upper bounds in Theorem 1 can be equal in applications of multi-stage stochastic linear optimization that arise in practice.

B.2.1. Example 1 from Section 3. Consider the multi-stage stochastic linear optimization problem

$$\begin{aligned} J^* = \quad & \underset{x_2:\mathbb{R}\to\mathbb{R},\, x_3:\mathbb{R}^2\to\mathbb{R}}{\text{minimize}} && \mathbb{E}\left[x_2(\xi_1) + 2x_3(\xi_1,\xi_2)\right] && \\ & \text{subject to} && x_2(\xi_1) + x_3(\xi_1,\xi_2) \ge \xi_1 + \xi_2 && \text{a.s.} \\ & && x_2(\xi_1),\, x_3(\xi_1,\xi_2) \ge 0 && \text{a.s.,} \end{aligned} \tag{3}$$

where the random variables $\boldsymbol{\xi} = (\xi_1,\xi_2) \in \mathbb{R}^2$ denote the preorder and regular demand of a new product. We assume that this stochastic process satisfies Assumption 1 and is contained in $\Xi := \mathbb{R}^2_+$.

Proposition 1. *For Problem* (3), $\underline{J} = J^*$*. If there is an optimal* $x_2^*:\mathbb{R}\to\mathbb{R}$ *for Problem* (3) *which is continuous, then* $\bar{J} = J^*$*.*

Proof. Our proof is split into two parts:

- For any arbitrary $\tilde{S} \subseteq S$, we observe that the optimal cost of the optimization problem

$$\begin{aligned} & \text{minimize} && \mathbb{E}\left[(x_2(\xi_1) + 2x_3(\xi_1,\xi_2))\,\mathbb{I}\left\{\boldsymbol{\xi}\in\tilde{S}\right\}\right] && \\ & \text{subject to} && x_2(\zeta_1) + x_3(\zeta_1,\zeta_2) \ge \xi_1 + \xi_2 && \forall \boldsymbol{\zeta}\in\tilde{S} \\ & && x_2(\zeta_1),\, x_3(\zeta_1,\zeta_2) \ge 0 && \forall \boldsymbol{\zeta}\in\tilde{S} \end{aligned} \tag{EC.9}$$

  would not change if we added the constraints

$$\begin{aligned} 0 \le x_2(\xi_1) \le \sup_{\boldsymbol{\zeta}\in\tilde{S}} \{\zeta_1+\zeta_2\} \quad & \text{a.s.} \\ 0 \le x_3(\xi_1,\xi_2) \le \xi_1 + \xi_2 \quad & \text{a.s.} \end{aligned}$$

  Moreover, any feasible solution to Problem (EC.9) which satisfies the above constraints will also satisfy

$$0 \le x_2(\xi_1) + 2x_3(\xi_1,\xi_2) \le 3\max\left\{\sup_{\boldsymbol{\zeta}\in\tilde{S}}\{\zeta_1+\zeta_2\}, \xi_1+\xi_2\right\} \quad \text{a.s.}$$

  as well as satisfy the constraints in Problem (3). Since Assumption 1 holds, we conclude from Theorem EC.1 that $\underline{J} = J^*$ for Problem (3).

- Let $x_2^*:\mathbb{R}\to\mathbb{R}$ denote an optimal second-stage decision rule to Problem (3) which is continuous. For any $M \ge 0$, define the new decision rules

$$x_2^M(\zeta_1) := \max\{0, \min\{x_2^*(\zeta_1), M\}\}, \quad x_3^M(\zeta_1,\zeta_2) := \max\left\{0, \zeta_1+\zeta_2 - x_2^M(\zeta_1)\right\}. \tag{EC.10}$$

  We observe that the decision rules from (EC.10) satisfy the constraints of Problem (3). Moreover,

$$\begin{aligned} & \mathbb{E}\left[x_2^M(\xi_1) + 2x_3^M(\xi_1,\xi_2)\right] \\ &= \mathbb{E}\left[x_2^M(\xi_1)\right] + \mathbb{E}\left[2x_3^M(\xi_1,\xi_2)\mathbb{I}\left\{x_2^*(\xi_1)\le M\right\}\right] + \mathbb{E}\left[2x_3^M(\xi_1,\xi_2)\mathbb{I}\left\{x_2^*(\xi_1) > M\right\}\right] && \text{(EC.11)} \\ &\le \mathbb{E}\left[x_2^*(\xi_1)\right] + \mathbb{E}\left[2x_3^M(\xi_1,\xi_2)\mathbb{I}\left\{x_2^*(\xi_1)\le M\right\}\right] + \mathbb{E}\left[2x_3^M(\xi_1,\xi_2)\mathbb{I}\left\{x_2^*(\xi_1) > M\right\}\right] && \text{(EC.12)} \\ &= \mathbb{E}\left[x_2^*(\xi_1)\right] + \mathbb{E}\left[2x_3^*(\xi_1,\xi_2)\mathbb{I}\left\{x_2^*(\xi_1)\le M\right\}\right] + \mathbb{E}\left[2x_3^M(\xi_1,\xi_2)\mathbb{I}\left\{x_2^*(\xi_1) > M\right\}\right] && \text{(EC.13)} \\ &\le J^* + \mathbb{E}\left[2x_3^M(\xi_1,\xi_2)\mathbb{I}\left\{x_2^*(\xi_1) > M\right\}\right] && \text{(EC.14)} \\ &\le J^* + 2\mathbb{E}\left[(\xi_1+\xi_2)\mathbb{I}\left\{x_2^*(\xi_1) > M\right\}\right] && \text{(EC.15)} \\ &\le J^* + 2\sqrt{\mathbb{E}\left[(\xi_1+\xi_2)^2\right]}\sqrt{\mathbb{P}\left(x_2^*(\xi_1) > M\right)}. && \text{(EC.16)} \end{aligned}$$

Indeed, (EC.11) follows from the law of total probability; (EC.12) and (EC.13) follow from the definition of the decision rules from (EC.10); (EC.14) holds because the inequality $x_3^M(\xi_1,\xi_2) \ge 0$ holds almost surely; (EC.15) follows from the definition of the decision rule $x_3^M(\xi_1,\xi_2)$; (EC.16) follows from the Cauchy-Schwartz inequality.

Since $\lim_{M\to\infty} \mathbb{P}(x_2^*(\xi_1) > M) = 0$ and $\mathbb{E}[(\xi_1+\xi_2)^2] < \infty$ (Assumption 1), we have shown, for every arbitrary $\eta > 0$, that there exists a constant $M \equiv M^\eta \ge 0$ such that the decision rules from (EC.10) are an η-optimal solution to Problem (3). Moreover, we readily observe that the decision rules from (EC.10) satisfy the properties of Theorem EC.2. Indeed, for every $M \ge 0$:

— Property (EC.6) is clearly satisfied by the decision rules (x_2^M, x_3^M).

— Property (EC.7) is satisfied by the decision rules (x_2^M, x_3^M) because the inequalities $0 \le x_2^M(\zeta_1) + 2x_3^M(\zeta_1,\zeta_2) \le M + 2(\zeta_1+\zeta_2)$ hold for all $\boldsymbol{\zeta} \in \Xi$.

— Property (EC.8) is satisfied by the decision rules (x_2^M, x_3^M) because the optimal decision rules (x_2^*, x_3^*) are continuous functions, which implies that (x_2^M, x_3^M) are continuous functions as well over the domain $\boldsymbol{\zeta} \in \Xi$.

We thus conclude from Theorem EC.2 that $J^* = \bar{J}$ for Problem (3).

□

B.2.2. Example from Section 7. Consider the multi-stage stochastic linear optimization problem

$$
\begin{array}{rlll}
J^* = & \underset{\mathbf{Q}\ge\mathbf{0},\mathbf{z}\in\{0,1\}^R,\mathbf{v}}{\text{minimize}} & \mathbb{E}\left[c\left(Q_{10}+\sum_{r=1}^R Q_{1r}\right) + hQ_{10} + \sum_{r=1}^R v_r(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) + f\sum_{r=1}^R z_r(\boldsymbol{\xi}_1)\right] & \\
& \text{subject to} & \sum_{r=1}^R Q_{2r}(\boldsymbol{\xi}_1) \le Q_{10} & \text{a.s.} \\
& & v_r(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) \ge b(\xi_{2r}+\xi_{1r}-Q_{2r}(\boldsymbol{\xi}_1)-Q_{1r}) - hQ_{2r}(\boldsymbol{\xi}_1) & \forall r\in[R],\ \text{a.s.} \\
& & v_r(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) \ge h(Q_{1r}-\xi_{1r}-\xi_{2r}) & \forall r\in[R],\ \text{a.s.} \\
& & v_r(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) \ge b(\xi_{1r}-Q_{1r}) - h\xi_{2r} & \forall r\in[R],\ \text{a.s.} \\
& & z_r(\boldsymbol{\xi}_1)\mathcal{M} \ge Q_{2r}(\boldsymbol{\xi}_1) & \forall r\in[R],\ \text{a.s.,}
\end{array} \tag{7}
$$

where the random variables $\boldsymbol{\xi} = (\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) \in \mathbb{R}^{2R}$ denote the demands of a weekly magazine at different retailers. We assume that this stochastic process satisfies Assumption 1 and is contained in $\Xi := \mathbb{R}_+^{2R}$. For simplicity, we focus on the case where $f = 0$.

PROPOSITION EC.1. *For Problem* (7), $\underline{J} = J^*$. *Moreover, if there are optimal decision rules* $Q_{21}^*,\ldots,Q_{2R}^* : \mathbb{R}^R \to \mathbb{R}$ *which are continuous, then* $\bar{J} = J^*$.

Proof. Our proof is split into two parts:

- For any arbitrary $\tilde{S} \subseteq S$, we observe that the optimal cost of the optimization problem

$$\begin{aligned}
\underset{\mathbf{Q} \geq \mathbf{0}, \mathbf{z} \in \{0,1\}^R, \mathbf{v}}{\text{minimize}} \quad & \mathbb{E}\left[\left(c\left(Q_{10} + \sum_{r=1}^R Q_{1r}\right) + hQ_{10} + \sum_{r=1}^R v_r(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)\right) \mathbb{I}\left\{\boldsymbol{\xi} \in \tilde{S}\right\}\right] \\
\text{subject to} \quad & \sum_{r=1}^R Q_{2r}(\boldsymbol{\zeta}_1) \leq Q_{10} && \forall \boldsymbol{\zeta} \in \tilde{S} \\
& v_r(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) \geq b(\zeta_{2r} + \zeta_{1r} - Q_{2r}(\boldsymbol{\zeta}_1) - Q_{1r}) - hQ_{2r}(\boldsymbol{\zeta}_1) && \forall r \in [R],\ \forall \boldsymbol{\zeta} \in \tilde{S} \\
& v_r(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) \geq h(Q_{1r} - \zeta_{1r} - \zeta_{2r}) && \forall r \in [R],\ \forall \boldsymbol{\zeta} \in \tilde{S} \\
& v_r(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) \geq b\,(\zeta_{1r} - Q_{1r}) - h\zeta_{2r} && \forall r \in [R],\ \forall \boldsymbol{\zeta} \in \tilde{S}
\end{aligned} \tag{EC.17}$$

would not change if we added the following constraints:

$$\begin{aligned}
& Q_{10} \leq \sup_{\boldsymbol{\zeta} \in \tilde{S}} \left\{ \sum_{r=1}^R (\zeta_{1r} + \zeta_{2r}) \right\} \\
& Q_{1r} \leq \sup_{\boldsymbol{\zeta} \in \tilde{S}} \{\zeta_{1r} + \zeta_{2r}\} \quad \forall r \in [R] \\
& \sum_{r=1}^R Q_{2r}(\boldsymbol{\xi}_1) \leq Q_{10} && \text{a.s.} \\
& v_r(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \leq \max\{b(\xi_{2r} + \xi_{1r}), hQ_{1r}, b\xi_{1r} - h\xi_{2r}\} && \forall r \in [R],\ \text{a.s.}
\end{aligned}$$

Moreover, any feasible solution to Problem (EC.17) which satisfies the above constraints will satisfy

$$0 \leq c\left(Q_{10} + \sum_{r=1}^R Q_{1r}\right) + hQ_{10} + \sum_{r=1}^R v_r(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \quad \text{a.s.},$$

will satisfy

$$\begin{aligned}
& c\left(Q_{10} + \sum_{r=1}^R Q_{1r}\right) + hQ_{10} + \sum_{r=1}^R v_r(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) \\
& \leq c\left(\sup_{\boldsymbol{\zeta} \in \tilde{S}} \left\{\sum_{r=1}^R (\zeta_{1r} + \zeta_{2r})\right\} + \sum_{r=1}^R \sup_{\boldsymbol{\zeta} \in \tilde{S}} \{\zeta_{1r} + \zeta_{2r}\}\right) + h \sup_{\boldsymbol{\zeta} \in \tilde{S}} \left\{\sum_{r=1}^R (\zeta_{1r} + \zeta_{2r})\right\} \\
& \quad + \sum_{r=1}^R \max\left\{b(\xi_{2r} + \xi_{1r}), h \sup_{\boldsymbol{\zeta} \in \tilde{S}} \{\zeta_{1r} + \zeta_{2r}\}, b\xi_{1r} - h\xi_{2r}\right\} && \text{a.s.} \\
& \leq (2c + 2h + b) \max\left\{\sum_{r=1}^R (\xi_{1r} + \xi_{2r}), \sup_{\boldsymbol{\zeta} \in \tilde{S}} \left\{\sum_{r=1}^R (\zeta_{1r} + \zeta_{2r})\right\}\right\} && \text{a.s.},
\end{aligned}$$

and will satisfy the constraints of Problem (7). Therefore, it readily follows that the conditions of Theorem EC.1 are satisfied, which implies that $\underline{J} = J^*$ for Problem (7).

- Let $Q_{21}^*, \ldots, Q_{2R}^* : \mathbb{R}^R \to \mathbb{R}$ denote optimal decision rules for Problem (7) which are continuous, and let $Q_{10}^*, \ldots, Q_{1R}^*$ denote optimal first-stage decisions for Problem (7). We define the following new decision rules for all retailers $r \in [R]$ and all realizations $\boldsymbol{\zeta} = (\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) \in \Xi$:

$$\begin{aligned}
Q'_{2r}(\boldsymbol{\zeta}_1) &:= \max\left\{0, \min\left\{Q_{21}^*(\boldsymbol{\zeta}_1), Q_{10}^* - \sum_{s=1}^{r-1} Q'_{2s}(\boldsymbol{\zeta}_1)\right\}\right\} \\
v'_r(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) &:= \max\{b(\zeta_{2r} + \zeta_{1r} - Q'_{2r}(\boldsymbol{\zeta}_1) - Q_{1r}^*) - hQ'_{2r}(\boldsymbol{\zeta}_1), h(Q_{1r}^* - \zeta_{1r} - \zeta_{2r}), b\,(\zeta_{1r} - Q_{1r}^*) - h\zeta_{2r}\}.
\end{aligned}$$

We observe that $(Q_{10}^*, \ldots, Q_{1R}^*, Q'_{21}, \ldots, Q'_{2R}, v'_1, \ldots, v'_R)$ is an optimal solution to Problem (7) and satisfies the conditions of Theorem EC.2, which concludes our proof that $J^* = \bar{J}$ for Problem (7).

□

B.2.3. Example from Section 8 Consider the multi-stage stochastic optimization problem

$$\begin{aligned}
J^* = \quad & \underset{\mathbf{x},\mathbf{I},\mathbf{y}}{\text{minimize}} && \mathbb{E}\left[\sum_{t=1}^T \left(c_t x_t(\xi_1,\ldots,\xi_{t-1}) + y_{t+1}(\xi_1,\ldots,\xi_t)\right)\right] \\
& \text{subject to} && I_{t+1}(\xi_1,\ldots,\xi_t) = I_t(\xi_1,\ldots,\xi_{t-1}) + x_t(\xi_1,\ldots,\xi_{t-1}) - \xi_t && \text{a.s., } \forall t \in [T] \\
& && y_{t+1}(\xi_1,\ldots,\xi_t) \ge h_t I_{t+1}(\xi_1,\ldots,\xi_t) && \text{a.s., } \forall t \in [T] \\
& && y_{t+1}(\xi_1,\ldots,\xi_t) \ge -b_t I_{t+1}(\xi_1,\ldots,\xi_t) && \text{a.s., } \forall t \in [T] \\
& && 0 \le x_t(\xi_1,\ldots,\xi_{t-1}) \le \bar{x}_t && \text{a.s., } \forall t \in [T].
\end{aligned} \tag{10}$$

where $I_1 = 0$ and the parameters $c_t, h_t, b_t, \bar{x}_t$ are nonnegative for all periods $t \in [T]$. We assume that the stochastic process $\boldsymbol{\xi} \equiv (\xi_1,\ldots,\xi_T) \in \mathbb{R}^T$ satisfies Assumption 1 and is contained in $\Xi = \mathbb{R}^T_+$.

PROPOSITION EC.2. *For Problem* (10), $\underline{J} = J^*$. *Moreover, if there exist optimal decision rules* $x_t^* : \mathbb{R}^{t-1} \to \mathbb{R}$ *which are continuous, then* $\bar{J} = J^*$.

Proof. Our proof is split into two parts:

- For any arbitrary $\tilde{S} \subseteq S$, we observe that the optimal cost of the optimization problem

$$\begin{aligned}
& \underset{\mathbf{x},\mathbf{I},\mathbf{y}}{\text{minimize}} && \mathbb{E}\left[\sum_{t=1}^T \left(c_t x_t(\xi_1,\ldots,\xi_{t-1}) + y_{t+1}(\xi_1,\ldots,\xi_t)\right) \mathbb{I}\left\{\boldsymbol{\xi} \in \tilde{S}\right\}\right] \\
& \text{subject to} && I_{t+1}(\zeta_1,\ldots,\zeta_t) = I_t(\zeta_1,\ldots,\zeta_{t-1}) + x_t(\zeta_1,\ldots,\zeta_{t-1}) - \zeta_t && \forall \boldsymbol{\zeta} \in \tilde{S},\ \forall t \in [T] \\
& && y_{t+1}(\zeta_1,\ldots,\zeta_t) \ge h_t I_{t+1}(\zeta_1,\ldots,\zeta_t) && \forall \boldsymbol{\zeta} \in \tilde{S},\ \forall t \in [T] \\
& && y_{t+1}(\zeta_1,\ldots,\zeta_t) \ge -b_t I_{t+1}(\zeta_1,\ldots,\zeta_t) && \forall \boldsymbol{\zeta} \in \tilde{S},\ \forall t \in [T] \\
& && 0 \le x_t(\zeta_1,\ldots,\zeta_{t-1}) \le \bar{x}_t && \forall \boldsymbol{\zeta} \in \tilde{S},\ \forall t \in [T].
\end{aligned} \tag{EC.18}$$

  would not change if we added the following constraints for each period $t \in [T]$:

$$\begin{aligned}
& I_{t+1}(\xi_1,\ldots,\xi_t) = I_t(\xi_1,\ldots,\xi_{t-1}) + x_t(\xi_1,\ldots,\xi_{t-1}) - \xi_t && \text{a.s.} \\
& y_{t+1}(\xi_1,\ldots,\xi_t) = \max\left\{h_t I_{t+1}(\xi_1,\ldots,\xi_t), -b_t I_{t+1}(\xi_1,\ldots,\xi_t)\right\} && \text{a.s.} \\
& 0 \le x_t(\xi_1,\ldots,\xi_{t-1}) \le \bar{x}_t && \text{a.s.}
\end{aligned}$$

  Moreover, any feasible solution to Problem (EC.18) which satisfies the above constraints is feasible for Problem (10) and also satisfies

$$0 \le \sum_{t=1}^T \left(c_t x_t(\xi_1,\ldots,\xi_{t-1}) + y_{t+1}(\xi_1,\ldots,\xi_t)\right) \le \left(\sum_{t=1}^T c_t + \sum_{t=1}^T h_t\right)\left(\sum_{t=1}^T \bar{x}_t\right) + \left(\sum_{t=1}^T b_t\right)\left(\sum_{t=1}^T \xi_t\right) \quad \text{a.s.}$$

  Therefore, it readily follows that the conditions of Theorem EC.1 are satisfied, which implies that $\underline{J} = J^*$ for Problem (10).

- Let $x_t^* : \mathbb{R}^{t-1} \to \mathbb{R}$ for each $t \in [T]$ denote optimal decision rules for Problem (10) which are continuous. We define the following new decision rules for each period $t \in [T]$ and all $\boldsymbol{\zeta} \in \Xi$:

$$\begin{aligned}
x_t'(\zeta_1,\ldots,\zeta_{t-1}) &:= \max\{0, \min\{x_t^*(\zeta_1,\ldots,\zeta_{t-1}), \bar{x}_t\}\} \\
I_{t+1}'(\zeta_1,\ldots,\zeta_t) &:= I_t'(\zeta_1,\ldots,\zeta_{t-1}) + x_t'(\zeta_1,\ldots,\zeta_{t-1}) - \zeta_t \\
y_{t+1}'(\zeta_1,\ldots,\zeta_t) &:= \max\left\{h_t I_{t+1}'(\zeta_1,\ldots,\zeta_t), -b_t I_{t+1}'(\zeta_1,\ldots,\zeta_t)\right\},
\end{aligned}$$

where $I_1' = 0$. We observe that $(x_1', \ldots, x_T', y_2', \ldots, y_{T+1}', I_1', \ldots, I_{T+1}')$ is an optimal solution to Problem (10) and satisfies the conditions of Theorem EC.2, which concludes our proof that $J^* = \bar{J}$ for Problem (10).

□

Appendix C: Proof of Theorem 1 from Section 4.2

In this appendix, we present the proof of Theorem 1. The theorem consists of asymptotic lower and upper bounds on the optimal cost of Problem (2), and we will address the proofs of the two bounds separately.

We first present the proof of the lower bound, which utilizes Theorem 2 from Section 4.2 and Theorem 3 from Section 4.4.

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

$$\underaccent{\tilde}{J} \le \liminf_{N \to \infty} \widehat{J}_N.$$

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

$$\begin{aligned}
\sum_{N=1}^{\infty} \mathbb{P}^N \left(\sup_{\boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j} L(1 + \|\boldsymbol{\zeta}\|) > \log N \right) &= \sum_{N=1}^{\infty} \mathbb{P}^N \left(\max_{j \in [N]} \left\{ L(1 + \|\hat{\boldsymbol{\xi}}^j\| + \epsilon_N) \right\} > \log N \right) && \text{(EC.19)} \\
&\le \sum_{N=1}^{\infty} N \mathbb{P} \left(L(1 + \|\boldsymbol{\xi}\| + \epsilon_N) > \log N \right) && \text{(EC.20)} \\
&= \sum_{N=1}^{\infty} N \mathbb{P} \left(\|\boldsymbol{\xi}\| > \frac{\log N}{L} - 1 - \epsilon_N \right) \\
&= \sum_{N=1}^{\infty} N \mathbb{P} \left(\exp(\|\boldsymbol{\xi}\|^a) > \exp\left(\left(\frac{\log N}{L} - 1 - \epsilon_N \right)^a \right) \right) \\
&\le \sum_{N=1}^{\infty} \frac{Nb}{\exp\left(\left(\frac{\log N}{L} - 1 - \epsilon_N \right)^a \right)} && \text{(EC.21)} \\
&< \infty, && \text{(EC.22)}
\end{aligned}$$

where (EC.19) follows from the definition of the uncertainty sets, (EC.20) follows from the union bound, (EC.21) follows from Markov's inequality, and (EC.22) follows from $a > 1$ and $\epsilon_N \to 0$. 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:

$$\begin{aligned}
\widehat{J}_N = \quad & \underset{\mathbf{x} \in \mathcal{X}}{\text{minimize}} && \frac{1}{N} \sum_{j=1}^{N} \sup_{\boldsymbol{\zeta} \in \mathcal{U}_N^j} \sum_{t=1}^{T} \mathbf{c}_t(\boldsymbol{\zeta}) \cdot \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \\
& \text{subject to} && \sum_{t=1}^{T} \mathbf{A}_t(\boldsymbol{\zeta}) \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \le \mathbf{b}(\boldsymbol{\zeta}) \qquad \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j \\
& && \left\| \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \right\| \le \log N \qquad \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j,\ t.
\end{aligned} \qquad \text{(EC.23)}$$

Moreover, since $\mathbf{c}_1(\boldsymbol{\zeta}) \in \mathbb{R}^{n_1}, \ldots, \mathbf{c}_T(\boldsymbol{\zeta}) \in \mathbb{R}^{n_T}$ are affine functions of the stochastic process, it follows from identical reasoning as (EC.19)-(EC.22) and the equivalence of ℓ_p-norms in finite-dimensional spaces that $\sup_{\boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j} \|\mathbf{c}_t(\boldsymbol{\zeta})\|_* \le \log N$ for all sufficiently large $N \in \mathbb{N}$, $\mathbb{P}^\infty$-almost surely.

We now apply Theorem 2 to obtain an asymptotic lower bound on the optimization problem in (EC.23). Indeed, let M_N be shorthand for $N^{-\frac{1}{(d+1)(d+2)}} \log N$. Then, for all sufficiently large $N \in \mathbb{N}$, $\mathbb{P}^\infty$-almost surely, and for any decision rule $\mathbf{x} \in \mathcal{X}$ which is feasible for the optimization problem in (EC.23),

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \mathbb{I}\left\{\boldsymbol{\xi} \in \cup_{j=1}^N \mathcal{U}_N^j\right\}\right] \\
&\le \frac{1}{N} \sum_{j=1}^N \sup_{\boldsymbol{\zeta} \in \mathcal{U}_N^j} \sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\zeta}) \cdot \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) + M_N \sup_{\boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j} \left| \sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\zeta}) \cdot \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \right| \\
&\le \frac{1}{N} \sum_{j=1}^N \sup_{\boldsymbol{\zeta} \in \mathcal{U}_N^j} \sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\zeta}) \cdot \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) + M_N \sum_{t=1}^T \sup_{\boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j} \|\mathbf{c}_t(\boldsymbol{\zeta})\|_* \|\mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1})\| \\
&\le \frac{1}{N} \sum_{j=1}^N \sup_{\boldsymbol{\zeta} \in \mathcal{U}_N^j} \sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\zeta}) \cdot \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) + T M_N (\log N)^2,
\end{aligned}
$$

where the first inequality follows from Theorem 2, the second inequality follows from the triangle inequality and the Cauchy-Schwartz inequality, and the third inequality follows because $\|\mathbf{c}_t(\boldsymbol{\zeta})\|_* \le \log N$ and $\|\mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T)\| \le \log N$ for all sufficiently large $N \in \mathbb{N}$ and all 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.23). Therefore, we have shown that the following inequality holds for all sufficiently large $N \in \mathbb{N}$, $\mathbb{P}^\infty$-almost surely:

$$
\begin{aligned}
\widehat{J}_N + T M_N (\log N)^2 \ge \quad & \underset{\mathbf{x} \in \mathcal{X}}{\text{minimize}} && \mathbb{E}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \mathbb{I}\left\{\boldsymbol{\xi} \in \cup_{j=1}^N \mathcal{U}_N^j\right\}\right] \\
& \text{subject to} && \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\zeta}) \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \le \mathbf{b}(\boldsymbol{\zeta}) && \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j \\
& && \left\|\mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1})\right\| \le \log N && \forall \boldsymbol{\zeta} \in \cup_{j=1}^N \mathcal{U}_N^j,\ t.
\end{aligned}
$$

Next, we obtain a lower bound on the right side of the above inequality by removing the last row of constraints and relaxing $\cup_{j=1}^N \mathcal{U}_N^j$ to any set which contains the stochastic process with sufficiently high probability:

$$
\begin{aligned}
& \underset{\mathbf{x} \in \mathcal{X}, \tilde{S} \subseteq \Xi}{\text{minimize}} && \mathbb{E}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \mathbb{I}\left\{\boldsymbol{\xi} \in \tilde{S}\right\}\right] \\
& \text{subject to} && \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\zeta}) \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \le \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \tilde{S} \\
& && \mathbb{P}\left(\boldsymbol{\xi} \in \tilde{S}\right) \ge \mathbb{P}\left(\boldsymbol{\xi} \in \cup_{j=1}^N \mathcal{U}_N^j\right).
\end{aligned}
$$

Finally, for any fixed $\rho \in (0,1)$, it follows from Theorem 3 that $\mathbb{P}(\boldsymbol{\xi} \in \cup_{j=1}^N \mathcal{U}_N^j \cap S) \ge 1 - \rho$ for all sufficiently large $N \in \mathbb{N}$, $\mathbb{P}^\infty$-almost surely.[3] Furthermore, we observe that $T M_N (\log N)^2$ converges to zero as the number

[3] We remark that Devroye and Wise (1980, Theorem 2) could be used here in lieu of Theorem 3. Our primary use of Theorem 3 is in the proof of Theorem 2.

of sample paths N tends to infinity. Therefore, we have shown that the following inequality holds, $\mathbb{P}^\infty$-almost surely:

$$\begin{aligned} \liminf_{N\to\infty} \widehat{J}_N \geq \quad \underset{\mathbf{x}\in\mathcal{X},\tilde{S}\subseteq\Xi}{\text{minimize}} \quad & \mathbb{E}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi})\cdot\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})\mathbb{I}\left\{\boldsymbol{\xi}\in\tilde{S}\right\}\right] \\ \text{subject to} \quad & \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\zeta})\mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \leq \mathbf{b}(\boldsymbol{\zeta}) \quad \forall\boldsymbol{\zeta}\in\tilde{S} \\ & \mathbb{P}\left(\boldsymbol{\xi}\in\tilde{S}\right) \geq 1-\rho. \end{aligned}$$

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

We now conclude the proof of Theorem 1 by establishing its upper bound.

THEOREM 1B. *Suppose Assumption 2 holds. Then,* $\mathbb{P}^\infty$*-almost surely we have*

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

Proof. Consider any $\rho>0$ such that there is a decision rule $\mathbf{x}\in\mathcal{X}$ which satisfies

$$\bar{\mathbb{E}}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi})\cdot\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})\right] < \infty, \tag{EC.24}$$

$$\sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\zeta})\mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \leq \mathbf{b}(\boldsymbol{\zeta}) \quad \forall\boldsymbol{\zeta}\in\Xi: \operatorname{dist}(\boldsymbol{\zeta},S)\leq\rho. \tag{EC.25}$$

Indeed, if no such $\rho>0$ and $\mathbf{x}\in\mathcal{X}$ 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,

$$\begin{aligned} \limsup_{N\to\infty}\widehat{J}_N &\leq \limsup_{N\to\infty}\frac{1}{N}\sum_{j=1}^N \sup_{\boldsymbol{\zeta}\in\Xi:\,\|\boldsymbol{\zeta}-\hat{\boldsymbol{\xi}}^j\|\leq\epsilon_N}\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\zeta})\cdot\mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \\ &\leq \lim_{k\to\infty}\limsup_{N\to\infty}\frac{1}{N}\sum_{j=1}^N \sup_{\boldsymbol{\zeta}\in\Xi:\,\|\boldsymbol{\zeta}-\hat{\boldsymbol{\xi}}^j\|\leq\epsilon_k}\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\zeta})\cdot\mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \\ &= \lim_{k\to\infty}\mathbb{E}\left[\sup_{\boldsymbol{\zeta}\in\Xi:\,\|\boldsymbol{\zeta}-\boldsymbol{\xi}\|\leq\epsilon_k}\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\zeta})\cdot\mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1})\right] \quad \mathbb{P}^\infty\text{-almost surely} \\ &= \bar{\mathbb{E}}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi})\cdot\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})\right]. \end{aligned} \tag{EC.26}$$

The first inequality holds because the decision rule is feasible but possibly suboptimal for Problem (2) for all $N\geq\min\{\bar{N}:\epsilon_{\bar{N}}\leq\rho\}$. The second inequality holds because $\epsilon_k\geq\epsilon_N$ for every fixed k and all $N\geq k$. The first equality follows from the strong law of large numbers (Erickson 1973), which holds since (EC.24) ensures that

$$\mathbb{E}\left[\max\left\{\sup_{\boldsymbol{\zeta}\in\Xi:\,\|\boldsymbol{\zeta}-\boldsymbol{\xi}\|\leq\epsilon_k}\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\zeta})\cdot\mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}),0\right\}\right] < \infty$$

for all sufficiently large k. The final equality follows the definition of the local upper semicontinuous envelope. Since the set of decision rules which satisfy (EC.25) does not get smaller as $\rho \downarrow 0$, we conclude that the following holds $\mathbb{P}^\infty$-almost surely:

$$\begin{aligned}\limsup_{N\to\infty} \widehat{J}_N \leq \quad \lim_{\rho \downarrow 0} \quad \underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \quad & \bar{\mathbb{E}}\left[\sum_{t=1}^T \mathbf{c}_t(\boldsymbol{\xi})\cdot \mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})\right]\\ \text{subject to} \quad & \sum_{t=1}^T \mathbf{A}_t(\boldsymbol{\zeta})\mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \leq \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta}\in\Xi:\ \mathrm{dist}(\boldsymbol{\zeta},S)\leq\rho.\end{aligned}$$

This concludes the proof. $\square$

Appendix D: Proof of Theorem 2 from Section 4.2

In this appendix, we present the proof of Theorem 2. The proof is organized as follows. In Appendix D.1, we first develop a helpful intermediary bound (Lemma EC.2). In Appendix D.2, we use that bound to prove Theorem 2. In Appendix D.3, we provide for completeness the proofs of some miscellaneous and rather technical results that were used in Appendix D.2.

D.1. An intermediary result

Our proof of Theorem 2 relies on an intermediary result (Lemma EC.2), which establishes a relationship between sample robust optimization and distributionally robust optimization with the 1-Wasserstein ambiguity set. We begin by establishing the relationship for the specific case where there is a single data point.

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

$$\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z}):\ \mathbb{E}_{\mathbb{Q}}[\|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}\|]\leq\theta_1} \mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{\xi})] \leq \sup_{\boldsymbol{\zeta}\in\mathcal{Z}:\ \|\boldsymbol{\zeta}-\hat{\boldsymbol{\xi}}\|\leq\theta_2} f(\boldsymbol{\zeta}) + \frac{2\theta_1}{\theta_2}\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}|f(\boldsymbol{\zeta})|. \tag{EC.27}$$

Proof. We first 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,

$$\begin{aligned}\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z}):\ \mathbb{E}_{\mathbb{Q}}[\|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}\|]\leq\theta_1} \mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{\xi})] &= \begin{cases}\displaystyle\sup_{\boldsymbol{\zeta}^1,\boldsymbol{\zeta}^2\in\mathcal{Z},\lambda\in[0,1]} & \lambda f(\boldsymbol{\zeta}^1)+(1-\lambda)f(\boldsymbol{\zeta}^2)\\ \text{subject to} & \lambda\left\|\boldsymbol{\zeta}^1-\hat{\boldsymbol{\xi}}\right\|+(1-\lambda)\left\|\boldsymbol{\zeta}^2-\hat{\boldsymbol{\xi}}\right\|\leq\theta_1\end{cases}\\ &\leq \begin{cases}\displaystyle\sup_{\boldsymbol{\zeta}^1,\boldsymbol{\zeta}^2\in\mathcal{Z},\lambda\in[0,1]} & \lambda f(\boldsymbol{\zeta}^1)+(1-\lambda)f(\boldsymbol{\zeta}^2)\\ \text{subject to} & \lambda\left\|\boldsymbol{\zeta}^1-\hat{\boldsymbol{\xi}}\right\|\leq\theta_1,\ (1-\lambda)\left\|\boldsymbol{\zeta}^2-\hat{\boldsymbol{\xi}}\right\|\leq\theta_1,\end{cases}\end{aligned} \tag{EC.28}$$

where the inequality follows from relaxing the constraints on $\boldsymbol{\zeta}^1$ and $\boldsymbol{\zeta}^2$. Let us assume from this point onward that $\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}|f(\boldsymbol{\zeta})|<\infty$; indeed, if $\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}|f(\boldsymbol{\zeta})|=\infty$, then the inequality in (EC.27) would trivially hold since the right-hand side would equal infinity. Then, it follows from (EC.28) that

$$\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z}):\ \mathbb{E}_{\mathbb{Q}}[\|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}\|]\leq\theta_1} \mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{\xi})] \leq \sup_{0\leq\lambda\leq1}\left\{\lambda\left(\sup_{\boldsymbol{\zeta}\in\mathcal{Z}:\ \|\boldsymbol{\zeta}-\hat{\boldsymbol{\xi}}\|\leq\frac{\theta_1}{\lambda}} f(\boldsymbol{\zeta})\right)+(1-\lambda)\left(\sup_{\boldsymbol{\zeta}\in\mathcal{Z}:\ \|\boldsymbol{\zeta}-\hat{\boldsymbol{\xi}}\|\leq\frac{\theta_1}{1-\lambda}} f(\boldsymbol{\zeta})\right)\right\}. \tag{EC.29}$$

We observe that the supremum over $0 \le \lambda \le 1$ is symmetric with respect to λ, in the sense that λ can be restricted to $[0, \frac{1}{2}]$ or $[\frac{1}{2}, 1]$ without loss of generality. Moreover, under the assumption that $\theta_2 \ge 2\theta_1$, 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.29) is equal to

$$\sup_{0 \le \lambda \le 1 - \frac{\theta_1}{\theta_2}} \left\{ \lambda \left(\sup_{\boldsymbol{\zeta} \in \mathcal{Z}:\, \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}\| \le \frac{\theta_1}{\lambda}} f(\boldsymbol{\zeta}) \right) + (1-\lambda) \left(\sup_{\boldsymbol{\zeta} \in \mathcal{Z}:\, \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}\| \le \frac{\theta_1}{1-\lambda}} f(\boldsymbol{\zeta}) \right) \right\}. \tag{EC.30}$$

Next, we observe that $\frac{\theta_1}{1-\lambda} \le \theta_2$ for every feasible λ for the above optimization problem. Using this inequality, we obtain the following upper bound:

$$\begin{aligned}
&\sup_{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}):\, \mathbb{E}_{\mathbb{Q}}[\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}\|] \le \theta_1} \mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{\xi})] \\
&\le \sup_{0 \le \lambda \le 1 - \frac{\theta_1}{\theta_2}} \left\{ \lambda \left(\sup_{\boldsymbol{\zeta} \in \mathcal{Z}:\, \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}\| \le \frac{\theta_1}{\lambda}} f(\boldsymbol{\zeta}) \right) + (1-\lambda) \left(\sup_{\boldsymbol{\zeta} \in \mathcal{Z}:\, \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}\| \le \theta_2} f(\boldsymbol{\zeta}) \right) \right\} \\
&= \sup_{\boldsymbol{\zeta} \in \mathcal{Z}:\, \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}\| \le \theta_2} f(\boldsymbol{\zeta}) + \sup_{0 \le \lambda \le 1 - \frac{\theta_1}{\theta_2}} \left\{ \lambda \left(\sup_{\boldsymbol{\zeta} \in \mathcal{Z}:\, \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}\| \le \frac{\theta_1}{\lambda}} f(\boldsymbol{\zeta}) - \sup_{\boldsymbol{\zeta} \in \mathcal{Z}:\, \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}\| \le \theta_2} f(\boldsymbol{\zeta}) \right) \right\},
\end{aligned} \tag{EC.31}$$

where the above equality comes from rearranging terms. For every $\frac{\theta_1}{\theta_2} \le \lambda \le 1 - \frac{\theta_1}{\theta_2}$, it immediately follows from $\frac{\theta_1}{\lambda} \le \theta_2$ that

$$\sup_{\boldsymbol{\zeta} \in \mathcal{Z}:\, \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}\| \le \frac{\theta_1}{\lambda}} f(\boldsymbol{\zeta}) - \sup_{\boldsymbol{\zeta} \in \mathcal{Z}:\, \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}\| \le \theta_2} f(\boldsymbol{\zeta}) \le 0,$$

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

$$\begin{aligned}
&\sup_{0 \le \lambda \le 1 - \frac{\theta_1}{\theta_2}} \left\{ \lambda \left(\sup_{\boldsymbol{\zeta} \in \mathcal{Z}:\, \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}\| \le \frac{\theta_1}{\lambda}} f(\boldsymbol{\zeta}) - \sup_{\boldsymbol{\zeta} \in \mathcal{Z}:\, \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}\| \le \theta_2} f(\boldsymbol{\zeta}) \right) \right\} \\
&= \sup_{0 \le \lambda \le \frac{\theta_1}{\theta_2}} \left\{ \lambda \left(\sup_{\boldsymbol{\zeta} \in \mathcal{Z}:\, \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}\| \le \frac{\theta_1}{\lambda}} f(\boldsymbol{\zeta}) - \sup_{\boldsymbol{\zeta} \in \mathcal{Z}:\, \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}\| \le \theta_2} f(\boldsymbol{\zeta}) \right) \right\} &\text{(EC.32)} \\
&\le \sup_{0 \le \lambda \le \frac{\theta_1}{\theta_2}} \left\{ \lambda \left(\sup_{\boldsymbol{\zeta} \in \mathcal{Z}} f(\boldsymbol{\zeta}) - \inf_{\boldsymbol{\zeta} \in \mathcal{Z}} f(\boldsymbol{\zeta}) \right) \right\} &\text{(EC.33)} \\
&\le \frac{2\theta_1}{\theta_2} \sup_{\boldsymbol{\zeta} \in \mathcal{Z}} |f(\boldsymbol{\zeta})|. &\text{(EC.34)}
\end{aligned}$$

Line (EC.32) follows because we can without loss of generality restrict λ to the interval $[0, \frac{\theta_1}{\theta_2}]$. Line (EC.33) is obtained by applying the global lower and upper bounds on $f(\boldsymbol{\zeta})$. Finally, we obtain (EC.34) since

$$0 \le \sup_{\boldsymbol{\zeta} \in \mathcal{Z}} f(\boldsymbol{\zeta}) - \inf_{\boldsymbol{\zeta} \in \mathcal{Z}} f(\boldsymbol{\zeta}) \le 2 \sup_{\boldsymbol{\zeta} \in \mathcal{Z}} |f(\boldsymbol{\zeta})|.$$

Combining (EC.31) and (EC.34), we obtain the desired result. □

We now extend the previous lemma to the general case with more than one data point. In the following, we let $\widehat{\mathbb{P}}_N$ denote the empirical distribution of historical data $\hat{\boldsymbol{\xi}}^1, \ldots, \hat{\boldsymbol{\xi}}^N \in \mathbb{R}^d$, $\mathcal{Z} \subseteq \mathbb{R}^d$ be any set that contains the historical data, and $\mathsf{d}_1(\cdot, \cdot)$ denote the 1-Wasserstein distance between two probability distributions (see Section 6).

Lemma EC.2. *Let $f:\mathbb{R}^d \to \mathbb{R}$ be measurable, $\mathcal{Z} \subseteq \mathbb{R}^d$, and $\hat{\boldsymbol{\xi}}^1,\dots,\hat{\boldsymbol{\xi}}^N \in \mathcal{Z}$. If $\theta_2 \ge 2\theta_1 \ge 0$, then*

$$\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z}):\, \mathsf{d}_1(\mathbb{Q},\widehat{\mathbb{P}}_N)\le\theta_1} \mathbb{E}_{\mathbb{Q}}[f(\boldsymbol{\xi})] \le \frac{1}{N}\sum_{j=1}^N \sup_{\boldsymbol{\zeta}\in\mathcal{Z}:\, \|\boldsymbol{\zeta}-\hat{\boldsymbol{\xi}}^j\|\le\theta_2} f(\boldsymbol{\zeta}) + \frac{4\theta_1}{\theta_2}\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}|f(\boldsymbol{\zeta})|.$$

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

$$\left\{\mathbb{Q}\in\mathcal{P}(\mathcal{Z}):\, \mathsf{d}_1(\mathbb{Q},\widehat{\mathbb{P}}_N)\le\theta_1\right\} = \left\{\frac{1}{N}\sum_{j=1}^N \mathbb{Q}_j :\, \begin{array}{r}\frac{1}{N}\sum_{j=1}^N \mathbb{E}_{\mathbb{Q}_j}\left[\|\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}^j\|\right]\le\theta_1\\ \mathbb{Q}_1,\dots,\mathbb{Q}_N\in\mathcal{P}(\mathcal{Z})\end{array}\right\}.$$

Therefore,

$$\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z}):\, \mathsf{d}_1(\mathbb{Q},\widehat{\mathbb{P}}_N)\le\theta_1} \mathbb{E}_{\mathbb{Q}}\left[f(\boldsymbol{\xi})\right] = \sup_{\boldsymbol{\gamma}\in\mathbb{R}^N_+}\left\{\frac{1}{N}\sum_{j=1}^N \sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z}):\, \mathbb{E}_{\mathbb{Q}}[\|\boldsymbol{\zeta}-\hat{\boldsymbol{\xi}}^j\|]\le\gamma_j} \mathbb{E}_{\mathbb{Q}}\left[f(\boldsymbol{\xi})\right]:\, \frac{1}{N}\sum_{j=1}^N\gamma_j\le\theta_1\right\}.$$

For any choice of $\boldsymbol{\gamma}\in\mathbb{R}^N_+$, we can partition the components γ_j into those that satisfy $2\gamma_j\le\theta_2$ and $2\gamma_j>\theta_2$. Thus,

$$\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z}):\, \mathsf{d}_1(\mathbb{Q},\widehat{\mathbb{P}}_N)\le\theta_1} \mathbb{E}_{\mathbb{Q}}\left[f(\boldsymbol{\xi})\right]\\
&\le \sup_{\boldsymbol{\gamma}\in\mathbb{R}^N_+}\left\{\frac{1}{N}\sum_{j\in[N]:\, 2\gamma_j\le\theta_2} \sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z}):\, \mathbb{E}_{\mathbb{Q}}[\|\boldsymbol{\zeta}-\hat{\boldsymbol{\xi}}^j\|]\le\gamma_j} \mathbb{E}_{\mathbb{Q}}\left[f(\boldsymbol{\xi})\right] + \frac{1}{N}\sum_{j\in[N]:\, 2\gamma_j>\theta_2}\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}|f(\boldsymbol{\zeta})|:\, \frac{1}{N}\sum_{j=1}^N\gamma_j\le\theta_1\right\}, \quad \text{(EC.35)}
\end{aligned}$$

where the inequality follows from upper bounding each of the inner distributionally robust optimization problems for which $2\gamma_j>\theta_2$ by $\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}|f(\boldsymbol{\zeta})|$. Due to the constraints on $\boldsymbol{\gamma}$, there can be at most $\frac{2N\theta_1}{\theta_2}$ components which satisfy $2\gamma_j>\theta_2$. It thus follows from (EC.35) that

$$\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z}):\, \mathsf{d}_1(\mathbb{Q},\widehat{\mathbb{P}}_N)\le\theta_1} \mathbb{E}_{\mathbb{Q}}\left[f(\boldsymbol{\xi})\right]\\
&\le \sup_{\boldsymbol{\gamma}\in\mathbb{R}^N_+}\left\{\frac{1}{N}\sum_{j\in[N]:\, 2\gamma_j\le\theta_2} \sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z}):\, \mathbb{E}_{\mathbb{Q}}[\|\boldsymbol{\zeta}-\hat{\boldsymbol{\xi}}^j\|]\le\gamma_j} \mathbb{E}_{\mathbb{Q}}\left[f(\boldsymbol{\xi})\right]:\, \frac{1}{N}\sum_{j=1}^N\gamma_j\le\theta_1\right\} + \frac{2\theta_1}{\theta_2}\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}|f(\boldsymbol{\zeta})|. \quad \text{(EC.36)}
\end{aligned}$$

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

$$\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z}):\, \mathsf{d}_1(\mathbb{Q},\widehat{\mathbb{P}}_N)\le\theta_1} \mathbb{E}_{\mathbb{Q}}\left[f(\boldsymbol{\xi})\right]\\
&\le \sup_{\boldsymbol{\gamma}\in\mathbb{R}^N_+}\left\{\frac{1}{N}\sum_{j\in[N]:\, 2\gamma_j\le\theta_2}\left(\sup_{\boldsymbol{\zeta}\in\mathcal{Z}:\, \|\boldsymbol{\zeta}-\hat{\boldsymbol{\xi}}^j\|\le\theta_2} f(\boldsymbol{\zeta}) + \frac{2\gamma_j}{\theta_2}\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}|f(\boldsymbol{\zeta})|\right): \frac{1}{N}\sum_{j=1}^N\gamma_j\le\theta_1\right\} + \frac{2\theta_1}{\theta_2}\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}|f(\boldsymbol{\zeta})| \quad \text{(EC.37)}\\
&\le \sup_{\boldsymbol{\gamma}\in\mathbb{R}^N_+}\left\{\frac{1}{N}\sum_{j=1}^N\left(\sup_{\boldsymbol{\zeta}\in\mathcal{Z}:\, \|\boldsymbol{\zeta}-\hat{\boldsymbol{\xi}}^j\|\le\theta_2} f(\boldsymbol{\zeta}) + \frac{2\gamma_j}{\theta_2}\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}|f(\boldsymbol{\zeta})|\right): \frac{1}{N}\sum_{j=1}^N\gamma_j\le\theta_1\right\} + \frac{2\theta_1}{\theta_2}\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}|f(\boldsymbol{\zeta})| \quad \text{(EC.38)}\\
&= \frac{1}{N}\sum_{j=1}^N \sup_{\boldsymbol{\zeta}\in\mathcal{Z}:\, \|\boldsymbol{\zeta}-\hat{\boldsymbol{\xi}}^j\|\le\theta_2} f(\boldsymbol{\zeta}) + \frac{4\theta_1}{\theta_2}\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}|f(\boldsymbol{\zeta})|. \quad \text{(EC.39)}
\end{aligned}$$

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

$$\sup_{\boldsymbol{\zeta}\in\mathcal{Z}:\, \|\boldsymbol{\zeta}-\hat{\boldsymbol{\xi}}^j\|\le\theta_2} f(\boldsymbol{\zeta}) + \frac{2\gamma_j}{\theta_2}\sup_{\boldsymbol{\zeta}\in\mathcal{Z}}|f(\boldsymbol{\zeta})| \ge 0$$

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 $\frac{1}{N}\sum_{j=1}^N\gamma_j\le\theta_1$. This concludes the proof. □

D.2. Proof of Theorem 2

We have established above a deterministic bound (Lemma EC.2) 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 prove Theorem 2. We remark that the following proof will employ Theorem 3 from Section 4.4 as well as notation from Section 6. For clarity of exposition, some intermediary and rather technical details of the following proof have been relegated to Appendix D.3.

THEOREM 2. *If Assumptions 1 and 2 hold, then there exists a $\bar{N} \in \mathbb{N}$, $\mathbb{P}^\infty$-almost surely, such that*

$$\mathbb{E}\left[f(\boldsymbol{\xi})\mathbb{I}\left\{\boldsymbol{\xi} \in \cup_{j=1}^N \mathcal{U}_N^j\right\}\right] \le \frac{1}{N}\sum_{j=1}^N \sup_{\boldsymbol{\zeta}\in\mathcal{U}_N^j} f(\boldsymbol{\zeta}) + M_N \sup_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j} |f(\boldsymbol{\zeta})|$$

for all $N \ge \bar{N}$ and all measurable functions $f : \mathbb{R}^d \to \mathbb{R}$, where $M_N := N^{-\frac{1}{(d+1)(d+2)}} \log N$.

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

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

It follows from Fournier and Guillin (2015) and Assumption 1 that $\mathsf{d}_1(\mathbb{P}, \widehat{\mathbb{P}}_N) \le \delta_N$ for all sufficiently large $N \in \mathbb{N}$, $\mathbb{P}^\infty$-almost surely (see Lemma EC.3 in Appendix D.3). Therefore, for every measurable function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\begin{aligned}
&\mathbb{E}\left[f(\boldsymbol{\xi})\mathbb{I}\left\{\boldsymbol{\xi} \in \cup_{j=1}^N \mathcal{U}_N^j\right\}\right] \\
&= \mathbb{E}\left[\left(f(\boldsymbol{\xi}) + \sup_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j} |f(\boldsymbol{\zeta})|\right)\mathbb{I}\left\{\boldsymbol{\xi} \in \cup_{j=1}^N \mathcal{U}_N^j\right\}\right] - \left(\sup_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j} |f(\boldsymbol{\zeta})|\right)\mathbb{P}(\boldsymbol{\xi} \in \cup_{j=1}^N \mathcal{U}_N^j) \\
&\le \sup_{\mathbb{Q}\in\mathcal{P}(\Xi):\, \mathsf{d}_1(\mathbb{Q},\widehat{\mathbb{P}}_N)\le\delta_N} \mathbb{E}_{\mathbb{Q}}\underbrace{\left[\left(f(\boldsymbol{\xi}) + \sup_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j} |f(\boldsymbol{\zeta})|\right)\mathbb{I}\left\{\boldsymbol{\xi} \in \cup_{j=1}^N \mathcal{U}_N^j\right\}\right]}_{g(\boldsymbol{\xi})} - \left(\sup_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j} |f(\boldsymbol{\zeta})|\right)\mathbb{P}(\boldsymbol{\xi} \in \cup_{j=1}^N \mathcal{U}_N^j),
\end{aligned} \tag{EC.40}$$

where the inequality holds for all sufficiently large $N \in \mathbb{N}$, $\mathbb{P}^\infty$-almost surely. Next, we observe that $g(\boldsymbol{\xi})$ equals zero when $\boldsymbol{\xi}$ is not an element of $\cup_{j=1}^N \mathcal{U}_N^j$ and is nonnegative otherwise. Therefore, without loss of generality, we can restrict the supremum over the expectation of $g(\boldsymbol{\xi})$ to distributions with support contained in $\cup_{j=1}^N \mathcal{U}_N^j$ (see Lemma EC.4 in Appendix D.3). Therefore, (EC.40) is equal to

$$\begin{aligned}
&\sup_{\mathbb{Q}\in\mathcal{P}(\cup_{j=1}^N\mathcal{U}_N^j):\, \mathsf{d}_1(\mathbb{Q},\widehat{\mathbb{P}}_N)\le\delta_N} \mathbb{E}_{\mathbb{Q}}\left[\left(f(\boldsymbol{\xi}) + \sup_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j} |f(\boldsymbol{\zeta})|\right)\mathbb{I}\left\{\boldsymbol{\xi} \in \cup_{j=1}^N \mathcal{U}_N^j\right\}\right] - \left(\sup_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j} |f(\boldsymbol{\zeta})|\right)\mathbb{P}(\boldsymbol{\xi} \in \cup_{j=1}^N \mathcal{U}_N^j) \\
&= \sup_{\mathbb{Q}\in\mathcal{P}(\cup_{j=1}^N\mathcal{U}_N^j):\, \mathsf{d}_1(\mathbb{Q},\widehat{\mathbb{P}}_N)\le\delta_N} \mathbb{E}_{\mathbb{Q}}\left[\left(f(\boldsymbol{\xi}) + \sup_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j} |f(\boldsymbol{\zeta})|\right)\right] - \left(\sup_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j} |f(\boldsymbol{\zeta})|\right)\mathbb{P}(\boldsymbol{\xi} \in \cup_{j=1}^N \mathcal{U}_N^j) \\
&= \sup_{\mathbb{Q}\in\mathcal{P}(\cup_{j=1}^N\mathcal{U}_N^j):\, \mathsf{d}_1(\mathbb{Q},\widehat{\mathbb{P}}_N)\le\delta_N} \mathbb{E}_{\mathbb{Q}}\left[f(\boldsymbol{\xi})\right] + \left(\sup_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j} |f(\boldsymbol{\zeta})|\right)\mathbb{P}(\boldsymbol{\xi} \notin \cup_{j=1}^N \mathcal{U}_N^j),
\end{aligned} \tag{EC.41}$$

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 $\cup_{j=1}^N \mathcal{U}_N^j$, and the second equality follows because $\sup_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j} |f(\boldsymbol{\zeta})|$ is independent of $\mathbb{Q}$. By Assumption 2 and the construction of δ_N, we have that $\epsilon_N \geq 2\delta_N$ for all sufficiently large $N\in\mathbb{N}$. Thus, it follows from Lemma EC.2 that (EC.41) is upper bounded by

$$\frac{1}{N}\sum_{j=1}^N \sup_{\boldsymbol{\zeta}\in\mathcal{U}_N^j} f(\boldsymbol{\zeta}) + \frac{4\delta_N}{\epsilon_N}\sup_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j}|f(\boldsymbol{\zeta})| + \left(\sup_{\boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j}|f(\boldsymbol{\zeta})|\right)\mathbb{P}(\boldsymbol{\xi}\notin\cup_{j=1}^N\mathcal{U}_N^j). \tag{EC.42}$$

By the definition of δ_N, and since $\epsilon_N = \kappa N^{-\frac{1}{3}}$ when $d=1$ and $\epsilon_N = \kappa N^{-\frac{1}{d+1}}$ when $d\geq 2$, we have that $\frac{4\delta_N}{\epsilon_N} \leq \frac{M_N}{2}$ for all sufficiently large N. Finally, Theorem 3 implies that $\mathbb{P}(\boldsymbol{\xi}\notin\cup_{j=1}^N\mathcal{U}_N^j) \leq \frac{M_N}{2}$ for all sufficiently large N, $\mathbb{P}^\infty$-almost surely. Combining (EC.40), (EC.41), and (EC.42), we obtain the desired result. □

D.3. Miscellaneous results

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

Lemma EC.3. *Suppose Assumption 1 holds, and let*

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

for any fixed $\bar\kappa>0$. Then, $\mathsf{d}_1(\mathbb{P},\widehat{\mathbb{P}}_N)\leq\delta_N$ 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 $b:=\mathbb{E}[\exp(\|\boldsymbol{\xi}\|^a)]<\infty$. Thus, it follows from Fournier and Guillin (2015, Theorem 2) that there exist constants $c_1,c_2>0$ (which depend only a, b, and d) such that for all $N\geq\bar N$,

$$\mathbb{P}^N\left(\mathsf{d}_1(\mathbb{P},\widehat{\mathbb{P}}_N)>\delta_N\right)\leq\begin{cases}c_1\exp\left(-c_2N\delta_N^2\right), & \text{if } d=1,\\ c_1\exp\left(-\frac{c_2N\delta_N^2}{(\log(2+1/\delta_N))^2}\right), & \text{if } d=2,\\ c_1\exp\left(-c_2N\delta_N^d\right), & \text{if } d\geq 3.\end{cases} \tag{EC.43}$$

First, suppose $d=1$ and $N\geq\bar N$. Then, it follows from the definition of $\delta_N=\bar\kappa N^{-\frac{1}{2}}\log N$ and (EC.43) that

$$\mathbb{P}^N\left(\mathsf{d}_1(\mathbb{P},\widehat{\mathbb{P}}_N)>\delta_N\right)\leq c_1\exp\left(-c_2N\delta_N^2\right)=c_1\exp\left(-c_2\bar\kappa^2(\log N)^2\right).$$

Second, suppose $d=2$ and $N\geq\bar N$. Then, it follows from the definition of $\delta_N=\bar\kappa N^{-\frac{1}{2}}(\log N)^2$ and (EC.43) that there exists some constant $\bar c>0$ (which depends only on $\bar\kappa$ and c_2) such that

$$\begin{aligned}\mathbb{P}^N\left(\mathsf{d}_1(\mathbb{P},\widehat{\mathbb{P}}_N)>\delta_N\right)&\leq c_1\exp\left(-\frac{c_2N\delta_N^2}{\log(2+1/\delta_N)^2}\right)\\&=c_1\exp\left(-\frac{c_2\bar\kappa^2(\log N)^4}{\log(2+\bar\kappa^{-2}N^{\frac{1}{2}}(\log N)^{-2})^2}\right)\\&\leq c_1\exp\left(-\frac{c_2\bar\kappa^2(\log N)^4}{\log(2+\bar\kappa^{-2}N^{\frac{1}{2}})^2}\right)\\&\leq c_1\exp\left(-\bar c(\log N)^2\right).\end{aligned}$$

Third, suppose $d \ge 3$ and $N \ge \bar{N}$. Then, it follows from the definition of $\delta_N = \bar{\kappa} N^{-\frac{1}{d}} (\log N)^2$ and (EC.43) that

$$\mathbb{P}^N \left(\mathsf{d}_1(\mathbb{P}, \widehat{\mathbb{P}}_N) > \delta_N \right) \le c_1 \exp\left(-c_2 N \delta_N^d\right) = c_1 \exp\left(-c_2 (\log N)^{2d}\right).$$

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

$$\sum_{N=1}^{\infty} \mathbb{P}^N \left(\mathsf{d}_1(\mathbb{P}, \widehat{\mathbb{P}}_N) > \delta_N \right) < \infty,$$

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

The second lemma (Lemma EC.4) 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.4. *Suppose $\Xi \subseteq \mathbb{R}^d$ and $\hat{\boldsymbol{\xi}}^1, \ldots, \hat{\boldsymbol{\xi}}^N \in \mathcal{Z} \subseteq \Xi$. Let $g : \Xi \to \mathbb{R}$ be any measurable function where $g(\boldsymbol{\zeta}) \ge 0$ for all $\boldsymbol{\zeta} \in \mathcal{Z}$. Then, for all $\theta \ge 0$,*

$$\sup_{\mathbb{Q} \in \mathcal{P}(\Xi):\, \mathsf{d}_1(\mathbb{Q}, \widehat{\mathbb{P}}_N) \le \theta} \mathbb{E}_{\mathbb{Q}}\left[g(\boldsymbol{\xi}) \mathbb{I}\{\boldsymbol{\xi} \in \mathcal{Z}\} \right] = \sup_{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}):\, \mathsf{d}_1(\mathbb{Q}, \widehat{\mathbb{P}}_N) \le \theta} \mathbb{E}_{\mathbb{Q}}\left[g(\boldsymbol{\xi}) \right].$$

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

$$\sup_{\mathbb{Q} \in \mathcal{P}(\Xi):\, \mathsf{d}_1(\mathbb{Q}, \widehat{\mathbb{P}}_N) \le \theta} \mathbb{E}_{\mathbb{Q}}\left[\bar{g}(\boldsymbol{\xi}) \right] \ge \sup_{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}):\, \mathsf{d}_1(\mathbb{Q}, \widehat{\mathbb{P}}_N) \le \theta} \mathbb{E}_{\mathbb{Q}}\left[\bar{g}(\boldsymbol{\xi}) \right] = \sup_{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}):\, \mathsf{d}_1(\mathbb{Q}, \widehat{\mathbb{P}}_N) \le \theta} \mathbb{E}_{\mathbb{Q}}\left[g(\boldsymbol{\xi}) \right],$$

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

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)),

$$\sup_{\mathbb{Q} \in \mathcal{P}(\Xi):\, \mathsf{d}_1(\mathbb{Q}, \widehat{\mathbb{P}}_N) \le \theta} \mathbb{E}_{\mathbb{Q}}\left[\bar{g}(\boldsymbol{\xi}) \right] = \begin{cases} \displaystyle\sup_{\boldsymbol{\zeta}^{j1}, \boldsymbol{\zeta}^{j2} \in \Xi,\, \lambda^j \in [0,1]} & \displaystyle\frac{1}{N} \sum_{j=1}^N \left(\lambda^j \bar{g}(\boldsymbol{\zeta}^{j1}) + (1 - \lambda^j) \bar{g}(\boldsymbol{\zeta}^{j2}) \right) \\ \text{subject to} & \displaystyle\frac{1}{N} \sum_{j=1}^N \left(\lambda^j \|\boldsymbol{\zeta}^{j1} - \hat{\boldsymbol{\xi}}^j\| + (1 - \lambda^j) \|\boldsymbol{\zeta}^{j2} - \hat{\boldsymbol{\xi}}^j\| \right) \le \theta. \end{cases}$$

For any arbitrary $\eta > 0$, let $(\bar{\boldsymbol{\zeta}}^{j1}, \bar{\boldsymbol{\zeta}}^{j2}, \bar{\lambda}^j)_{j \in [N]}$ be an η-optimal solution to the above optimization problem. We now perform a transformation on this solution. For each $j \in [N]$, define $\breve{\lambda}^j = \bar{\lambda}^j$, and for each $* \in \{1, 2\}$, define $\breve{\boldsymbol{\zeta}}^{j*} = \bar{\boldsymbol{\zeta}}^{j*}$ if $\bar{\boldsymbol{\zeta}}^{j*} \in \mathcal{Z}$ and $\breve{\boldsymbol{\zeta}}^{j*} = \hat{\boldsymbol{\xi}}^j$ otherwise. Since $\bar{g}(\boldsymbol{\zeta}) \ge 0$ for all $\boldsymbol{\zeta} \in \Xi$ and $\bar{g}(\boldsymbol{\zeta}) = 0$ for all $\boldsymbol{\zeta} \notin \mathcal{Z}$, it follows that $\bar{g}(\breve{\boldsymbol{\zeta}}^{j*}) \ge \bar{g}(\bar{\boldsymbol{\zeta}}^{j*})$. By construction, $(\breve{\boldsymbol{\zeta}}^{j1}, \breve{\boldsymbol{\zeta}}^{j2}, \breve{\lambda}^j)_{j \in [N]}$ is a feasible solution to the above optimization problem, and is also feasible for

$$\begin{aligned} &\sup_{\boldsymbol{\zeta}^{j1}, \boldsymbol{\zeta}^{j2} \in \mathcal{Z},\, \lambda^j \in [0,1]} && \frac{1}{N} \sum_{j=1}^N \left(\lambda^j \bar{g}(\boldsymbol{\zeta}^{j1}) + (1 - \lambda^j) \bar{g}(\boldsymbol{\zeta}^{j2}) \right) \\ &\text{subject to} && \frac{1}{N} \sum_{j=1}^N \left(\lambda^j \|\boldsymbol{\zeta}^{j1} - \hat{\boldsymbol{\xi}}^j\| + (1 - \lambda^j) \|\boldsymbol{\zeta}^{j2} - \hat{\boldsymbol{\xi}}^j\| \right) \le \theta, \end{aligned}$$

where we replaced the domain of $\boldsymbol{\zeta}^{j1}$ and $\boldsymbol{\zeta}^{j2}$ by $\mathcal{Z}$. We have thus shown that

$$\begin{aligned}\sup_{\mathbb{Q}\in\mathcal{P}(\Xi):\, \mathsf{d}_1(\mathbb{P},\widehat{\mathbb{P}}_N)\le\theta} \mathbb{E}_{\mathbb{Q}}\left[\bar{g}(\boldsymbol{\xi})\right] &\le \frac{1}{N}\sum_{j=1}^{N}\left(\bar{\lambda}^j\bar{g}(\bar{\boldsymbol{\zeta}}^{j1})+(1-\bar{\lambda}^j)\bar{g}(\bar{\boldsymbol{\zeta}}^{j2})\right)+\eta \\ &\le \frac{1}{N}\sum_{j=1}^{N}\left(\breve{\lambda}^j\bar{g}(\breve{\boldsymbol{\zeta}}^{j1})+(1-\breve{\lambda}^j)\bar{g}(\breve{\boldsymbol{\zeta}}^{j2})\right)+\eta \le \sup_{\mathbb{Q}\in\mathcal{P}(\mathcal{Z}):\, \mathsf{d}_1(\mathbb{Q},\widehat{\mathbb{P}}_N)\le\theta} \mathbb{E}_{\mathbb{Q}}\left[\bar{g}(\boldsymbol{\xi})\right]+\eta.\end{aligned}$$

Since $\eta>0$ was chosen arbitrarily, and by the equivalence of $\bar{g}$ and g on $\mathcal{Z}$, we have shown the other direction. This concludes the proof. $\square$

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 below.

Consider the single-stage stochastic problem

$$\begin{aligned}&\underset{x_1\in\mathbb{Z}}{\text{minimize}} && x_1\\ &\text{subject to} && x_1\ge\xi_1 \quad \text{a.s.,}\end{aligned}$$

where the random variable ξ_1 is governed by the probability distribution $\mathbb{P}(\xi_1>\alpha)=(1-\alpha)^k$ for fixed $k>0$, and $\Xi=[0,2]$. We observe that the support of the random variable is $S=[0,1]$, and thus the optimal cost of the stochastic problem is $J^*=1$. We similarly observe that the lower bound is $\underline{J}=1$ and the upper bound, due to the integrality of the first stage decision, is $\bar{J}=2$. If $\epsilon_N=N^{-\frac{1}{3}}$, then we prove in Appendix E that the bounds in Theorem 1 are tight under different choices of k:

Range of k	*Result*
$k\in(0,3)$	$\mathbb{P}^\infty\left(\underline{J}<\liminf_{N\to\infty}\widehat{J}_N=\limsup_{N\to\infty}\widehat{J}_N=\bar{J}\right)=1$
$k=3$	$\mathbb{P}^\infty\left(\underline{J}=\liminf_{N\to\infty}\widehat{J}_N<\limsup_{N\to\infty}\widehat{J}_N=\bar{J}\right)=1$
$k\in(3,\infty)$	$\mathbb{P}^\infty\left(\underline{J}=\liminf_{N\to\infty}\widehat{J}_N=\limsup_{N\to\infty}\widehat{J}_N<\bar{J}\right)=1$

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

$$\begin{aligned}\underline{J}&=\lim_{\rho\downarrow0}\min_{x_1\in\mathbb{Z}}\left\{x_1:\mathbb{P}(x_1\ge\xi_1)\ge1-\rho\right\}=1, \text{ and}\\ \bar{J}&=\lim_{\rho\downarrow0}\min_{x_1\in\mathbb{Z}}\left\{x_1:x_1\ge1+\rho\right\}=2.\end{aligned}$$

Furthermore, given historical data, the choice of the robustness parameter $\epsilon_N=N^{-\frac{1}{3}}$, and $\Xi=[0,2]$,

$$\widehat{J}_N=\min_{x_1\in\mathbb{Z}}\left\{x_1:x_1\ge\zeta_1,\,\forall\zeta_1\in\cup_{j=1}^N\mathcal{U}_N^j\right\}=\begin{cases}1, & \text{if } \max_{j\in[N]}\hat{\xi}_1^j\le1-N^{-\frac{1}{3}},\\ 2, & \text{if } \max_{j\in[N]}\hat{\xi}_1^j>1-N^{-\frac{1}{3}}.\end{cases}$$

We first show that

$$\mathbb{P}^\infty\left(\limsup_{N\to\infty}\widehat{J}_N=1\right)=\begin{cases}0, & \text{if } 0<k\le3,\\ 1, & \text{if } k>3.\end{cases} \qquad \text{(Claim 1)}$$

Indeed,

$$
\begin{aligned}
&\mathbb{P}^\infty\left(\limsup_{N\to\infty} \widehat{J}_N = 1\right)\\
&= \mathbb{P}^\infty\left(\max_{j\in[N]} \hat{\xi}_1^j \le 1 - N^{-\frac{1}{3}} \text{ for all sufficiently large } N\right)\\
&= \lim_{N\to\infty} \mathbb{P}^\infty\left(\max_{j\in[n]} \hat{\xi}_1^j \le 1 - n^{-\frac{1}{3}} \text{ for all } n \ge N\right)\\
&= \lim_{N\to\infty} \mathbb{P}^\infty\left(\max_{j\in[N]} \hat{\xi}_1^j \le 1 - N^{-\frac{1}{3}} \text{ and } \max_{j\in[n]} \hat{\xi}_1^j \le 1 - n^{-\frac{1}{3}} \text{ for all } n \ge N+1\right)\\
&= \lim_{N\to\infty} \mathbb{P}^N\left(\max_{j\in[N]} \hat{\xi}_1^j \le 1 - N^{-\frac{1}{3}}\right) \prod_{n=N+1}^{\infty} \mathbb{P}\left(\max_{j\in[n]} \hat{\xi}_1^j \le 1 - n^{-\frac{1}{3}} \,\middle|\, \max_{j\in[n-1]} \hat{\xi}_1^j \le 1 - (n-1)^{-\frac{1}{3}}\right) &\text{(EC.44)}\\
&= \lim_{N\to\infty} \mathbb{P}^\infty\left(\max_{j\in[N]} \hat{\xi}_1^j \le 1 - N^{-\frac{1}{3}}\right) \prod_{n=N+1}^{\infty} \mathbb{P}\left(\hat{\xi}_1^n \le 1 - n^{-\frac{1}{3}} \,\middle|\, \max_{j\in[n-1]} \hat{\xi}_1^j \le 1 - (n-1)^{-\frac{1}{3}}\right) &\text{(EC.45)}\\
&= \lim_{N\to\infty} \mathbb{P}\left(\xi_1 \le 1 - N^{-\frac{1}{3}}\right)^N \prod_{n=N+1}^{\infty} \mathbb{P}\left(\xi_1 \le 1 - n^{-\frac{1}{3}}\right) &\text{(EC.46)}\\
&= \lim_{N\to\infty} \left(1 - N^{-\frac{k}{3}}\right)^N \prod_{n=N+1}^{\infty} \left(1 - n^{-\frac{k}{3}}\right). &\text{(EC.47)}
\end{aligned}
$$

Line (EC.44) follows from the law of total probability. Line (EC.45) follows because, conditional on $\max_{j\in[n-1]} \hat{\xi}_1^j \le 1 - (n-1)^{-\frac{1}{3}}$, we have that $\hat{\xi}_1^j \le 1 - n^{-\frac{1}{3}}$ for all $j \in [n-1]$. Line (EC.46) follows from the independence of $\hat{\xi}^j$, $j \in \mathbb{N}$. By evaluating the limit in (EC.47), we conclude the proof of Claim 1.

Next, we show that

$$\mathbb{P}^\infty\left(\liminf_{N\to\infty} \widehat{J}_N = 1\right) = 1 \text{ if } k \ge 3. \tag{Claim 2}$$

Indeed,

$$
\begin{aligned}
\mathbb{P}^\infty\left(\liminf_{N\to\infty} \widehat{J}_N = 1\right) &= \mathbb{P}^\infty\left(\max_{j\in[N]} \hat{\xi}_1^j \le 1 - N^{-\frac{1}{3}} \text{ for infinitely many } N\right)\\
&= \lim_{N\to\infty} \mathbb{P}^\infty\left(\max_{j\in[n]} \hat{\xi}_1^j \le 1 - n^{-\frac{1}{3}} \text{ for some } n \ge N\right)\\
&\ge \lim_{N\to\infty} \mathbb{P}^N\left(\max_{j\in[N]} \hat{\xi}_1^j \le 1 - N^{-\frac{1}{3}}\right)\\
&= \lim_{N\to\infty} \mathbb{P}\left(\xi_1 \le 1 - N^{-\frac{1}{3}}\right)^N &\text{(EC.48)}\\
&= \lim_{N\to\infty} \left(1 - N^{-\frac{k}{3}}\right)^N. &\text{(EC.49)}
\end{aligned}
$$

Line (EC.48) follows from the independence of $\hat{\xi}^j$, $j \in \mathbb{N}$. We observe that the limit in (EC.49) is strictly positive when $k \ge 3$. It follows from the Hewitt-Savage zero-one law (see, *e.g.*, Breiman (1992), Wang and Tomkins (1992)) that the event $\{\max_{j\in[N]} \hat{\xi}_1^j \le 1 - N^{-\frac{1}{3}}$ for infinitely many $N\}$ happens with probability zero or one. Thus, (EC.49) implies that the event $\{\liminf_{N\to\infty} \widehat{J}_N = 1\}$ must occur with probability one for $k \ge 3$.

Finally, we show that

$$\mathbb{P}^\infty\left(\liminf_{N\to\infty} \widehat{J}_N = 1\right) = 0 \text{ if } 0 < k < 3. \tag{Claim 3}$$

Indeed, suppose that $0 < k < 3$. Then,

$$\sum_{N=1}^{\infty} \mathbb{P}^{\infty}\left(\widehat{J}_N = 1\right) = \sum_{N=1}^{\infty} \mathbb{P}^{N}\left(\max_{j \in [N]} \hat{\xi}_1^j \le 1 - N^{-\frac{1}{3}}\right) = \sum_{N=1}^{\infty}\left(1 - N^{-\frac{k}{3}}\right)^N < \infty.$$

Therefore, it follows from the Borel-Cantelli lemma that

$$\mathbb{P}^{\infty}\left(\liminf_{N\to\infty} \widehat{J}_N = 1\right) = \mathbb{P}^{\infty}\left(\max_{j\in[N]} \hat{\xi}_1^j > 1 - N^{-\frac{1}{3}} \text{ for all sufficiently large } N\right) = 0,$$

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

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

Appendix F: Proof of Theorem 3 from Section 4.4

In this appendix, we present the proof of Theorem 3. Our proof techniques follow similar reasoning to Devroye and Wise (1980) and Baíllo et al. (2000) for $S_N := \cup_{j=1}^N \mathcal{U}_N^j$, which we adapt to Assumption 1. We remark that the following theorem also provides an intermediary step in the proofs of Theorems 1 and 2, which are found in Appendices C and D.

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

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

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

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

We begin by showing that $R_N\mathbb{P}(\boldsymbol{\xi} \notin A_N) \le \eta$ for all sufficiently large $N \in \mathbb{N}$. Indeed,

$$R_N\mathbb{P}(\boldsymbol{\xi} \notin A_N) = R_N\mathbb{P}\left(\|\boldsymbol{\xi}\| > (\log N)^{\frac{a+1}{2a}}\right) = R_N\mathbb{P}\left(\exp(\|\boldsymbol{\xi}\|^a) > \exp((\log N)^{\frac{a+1}{2}})\right) \le \frac{bR_N}{\exp((\log N)^{\frac{a+1}{2}})} \le \eta.$$

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

$$\alpha_N := \frac{\eta}{(\log N)^{\frac{d(a+1)}{2a}}\phi R_N}, \qquad B_N := \left\{\boldsymbol{\zeta} \in \mathbb{R}^d : \mathbb{P}(\|\boldsymbol{\xi} - \boldsymbol{\zeta}\| \le \epsilon_N) > \alpha_N \epsilon_N^d\right\},$$

where $\phi > 0$ is a constant which depends only on d and will be defined shortly. We now show that $R_N\mathbb{P}(\boldsymbol{\xi} \notin B_N) \le 2\eta$ for all sufficiently large N. Indeed, for all sufficiently large $N \in \mathbb{N}$,

$$\begin{aligned} R_N\mathbb{P}(\boldsymbol{\xi} \notin B_N) &= R_N\mathbb{P}(\boldsymbol{\xi} \in A_N, \boldsymbol{\xi} \notin B_N) + R_N\mathbb{P}(\boldsymbol{\xi} \notin A_N, \boldsymbol{\xi} \notin B_N) \\ &\le R_N\mathbb{P}(\boldsymbol{\xi} \in A_N, \boldsymbol{\xi} \notin B_N) + R_N\mathbb{P}(\boldsymbol{\xi} \notin A_N) \\ &\le R_N\mathbb{P}(\boldsymbol{\xi} \in A_N, \boldsymbol{\xi} \notin B_N) + \eta, \end{aligned} \tag{EC.50}$$

where the final inequality follows because $R_N\mathbb{P}(\boldsymbol{\xi} \notin A_N) \le \eta$ for all sufficiently large $N \in \mathbb{N}$. Now, choose points $\boldsymbol{\zeta}^1, \ldots, \boldsymbol{\zeta}^{K_N} \in A_N$ such that $\min_{j\in[K_N]} \|\boldsymbol{\zeta} - \boldsymbol{\zeta}^j\| \le \frac{\epsilon_N}{2}$ for all $\boldsymbol{\zeta} \in A_N$. 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

$$K_N \le \phi \left(\frac{(\log N)^{\frac{a+1}{2a}}}{\epsilon_N} \right)^d, \tag{EC.51}$$

where $\phi > 0$ is a constant that depends only on d. Then, continuing from (EC.50),

$$R_N \mathbb{P}(\boldsymbol{\xi} \notin B_N) \le R_N \mathbb{P}(\boldsymbol{\xi} \in A_N, \boldsymbol{\xi} \notin B_N) + \eta \le R_N \sum_{j=1}^{K_N} \mathbb{P}\left(\|\boldsymbol{\xi} - \boldsymbol{\zeta}^j\| \le \frac{\epsilon_N}{2}, \boldsymbol{\xi} \notin B_N \right) + \eta, \tag{EC.52}$$

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 $\boldsymbol{\zeta} \notin B_N$ such that $\|\boldsymbol{\zeta} - \boldsymbol{\zeta}^j\| \le \frac{\epsilon_N}{2}$. Then,

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

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

$$\mathbb{P}\left(\|\boldsymbol{\xi} - \boldsymbol{\zeta}^j\| \le \frac{\epsilon_N}{2}, \boldsymbol{\xi} \notin B_N \right) = 0.$$

In each of the two cases, we have shown that

$$\mathbb{P}\left(\|\boldsymbol{\xi} - \boldsymbol{\zeta}^j\| \le \frac{\epsilon_N}{2}, \boldsymbol{\xi} \notin B_N \right) \le \alpha_N \epsilon_N^d \tag{EC.53}$$

for each $j \in [K_N]$. Therefore, we combine (EC.52) and (EC.53) to obtain the following upper bound on $R_N \mathbb{P}(\boldsymbol{\xi} \notin B_N)$ for all sufficiently large $N \in \mathbb{N}$:

$$R_N \mathbb{P}(\boldsymbol{\xi} \notin B_N) \le R_N K_N \alpha_N \epsilon_N^d + \eta \le (\log N)^{\frac{d(a+1)}{2a}} \phi R_N \alpha_N + \eta \le 2\eta. \tag{EC.54}$$

The first inequality follows from (EC.52) and (EC.53), the second inequality follows from (EC.51), and the third inequality follows from the definition of α_N.

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

$$R_N \mathbb{P}(\boldsymbol{\xi} \notin S_N) = R_N \mathbb{P}(\boldsymbol{\xi} \notin S_N, \boldsymbol{\xi} \in B_N) + R_N \mathbb{P}(\boldsymbol{\xi} \notin S_N, \boldsymbol{\xi} \notin B_N) \le R_N \mathbb{P}(\boldsymbol{\xi} \notin S_N, \boldsymbol{\xi} \in B_N) + 2\eta, \tag{EC.55}$$

where the equality follows from the law of total probability and the inequality follows from (EC.54). Let $\rho := \frac{d(a-1)}{2a} > 0$. Then, for all sufficiently large $N \in \mathbb{N}$:

$$\begin{aligned}
\mathbb{P}^N \left(R_N \mathbb{P}(\boldsymbol{\xi} \notin S_N) > 3\eta \right) &\le \mathbb{P}^N \left(R_N \mathbb{P}(\boldsymbol{\xi} \notin S_N, \boldsymbol{\xi} \in B_N) > \eta \right) && \text{(EC.56)}\\
&\le \eta^{-1} R_N \mathbb{E}_{\mathbb{P}^N} \left[\mathbb{P}\left(\boldsymbol{\xi} \notin S_N, \boldsymbol{\xi} \in B_N \right) \right] && \text{(EC.57)}\\
&= \eta^{-1} R_N \mathbb{E}\left[\mathbb{I}\{\boldsymbol{\xi} \in B_N\} \mathbb{P}^N \left(\boldsymbol{\xi} \notin S_N \right) \right] && \text{(EC.58)}\\
&= \eta^{-1} R_N \mathbb{E}\left[\mathbb{I}\{\boldsymbol{\xi} \in B_N\} \mathbb{P}^N \left(\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^1\| > \epsilon_N, \ldots, \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^N\| > \epsilon_N \right) \right] && \text{(EC.59)}\\
&= \eta^{-1} R_N \mathbb{E}\left[\mathbb{I}\{\boldsymbol{\xi} \in B_N\} \mathbb{P}^1 \left(\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^1\| > \epsilon_N \right)^N \right] && \text{(EC.60)}\\
&\le \eta^{-1} R_N (1 - \alpha_N \epsilon_N^d)^N && \text{(EC.61)}\\
&\le \eta^{-1} R_N \exp\left(-N \alpha_N \epsilon_N^d \right) && \text{(EC.62)}\\
&\le \eta^{-1} R_N \exp\left(-\kappa^d N^{\frac{1}{d+1}} \alpha_N \right) && \text{(EC.63)}\\
&= \eta^{-1} R_N \exp\left(-\kappa^d \eta \phi^{-1} (\log N)^{1+\rho} \right). && \text{(EC.64)}
\end{aligned}$$

Line (EC.56) follows from (EC.55), (EC.57) follows from Markov's inequality, (EC.58) follows from Fubini's thoerem, and (EC.59) follows from the definition of S_N. Line (EC.60) follows because, given any fixed realization of $\boldsymbol{\xi}$, the random variables $\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^1\|, \ldots, \|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^N\|$ are independent. Note that the random vector $\hat{\boldsymbol{\xi}}^1$ is drawn from the measure $\mathbb{P}^1$. Line (EC.61) follows from the definition of B_N and the fact that the function $\boldsymbol{\xi} \mapsto \mathbb{I}\{\boldsymbol{\xi} \in B_N\}\, \mathbb{P}^1\left(\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^1\| > \epsilon_N\right)^N$ is equal to zero if $\boldsymbol{\xi} \notin B_N$. Line (EC.62) follows from the mean value theorem. Line (EC.63) holds since Assumption 2 implies that $\epsilon_N \ge \kappa N^{-\frac{1}{d+1}}$. Line (EC.64) follows from the definitions of α_N, R_N, and ρ. Since $\rho > 0$, it follows from (EC.64) and the definition of R_N that

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

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

Appendix G: Proof of Proposition 3 from Section 6

In this appendix, we present the proof of Proposition 3. We begin with the following intermediary lemma.

Lemma EC.5. *The ∞-Wasserstein ambiguity set is equivalent to*

$$\left\{ \frac{1}{N}\sum_{j=1}^N \mathbb{Q}_j : \begin{array}{l} \mathbb{Q}_j\left(\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^j\| \le \epsilon_N\right) = 1 \quad \forall j \in [N], \\ \mathbb{Q}_1, \ldots, \mathbb{Q}_N \in \mathcal{P}(\Xi) \end{array} \right\}.$$

Proof. By the definition of the ∞-Wasserstein distance from Section 6,

$$\left\{\mathbb{Q} \in \mathcal{P}(\Xi) : \mathsf{d}_\infty\left(\mathbb{Q}, \widehat{\mathbb{P}}_N\right) \le \epsilon_N\right\} = \left\{ \mathbb{Q} \in \mathcal{P}(\Xi) : \begin{array}{l} \Pi \in \mathcal{P}(\Xi \times \Xi), \\ \Pi\left(\|\boldsymbol{\xi} - \boldsymbol{\xi}'\| \le \epsilon_N\right) = 1, \text{ and} \\ \Pi \text{ is a joint distribution of } \boldsymbol{\xi} \text{ and } \boldsymbol{\xi}' \\ \text{with marginals } \mathbb{Q} \text{ and } \widehat{\mathbb{P}}_N, \text{ respectively} \end{array} \right\}. \tag{EC.65}$$

Let $\bar{\boldsymbol{\xi}}^1, \ldots, \bar{\boldsymbol{\xi}}^L$ be the distinct vectors among $\hat{\boldsymbol{\xi}}^1, \ldots, \hat{\boldsymbol{\xi}}^N$, and let $I_1, \ldots, I_L$ be index sets defined as

$$I_\ell := \{j \in [N] : \hat{\boldsymbol{\xi}}^j = \bar{\boldsymbol{\xi}}^\ell\}.$$

For any joint distribution Π that satisfies the constraints in the ambiguity set in (EC.65), let $\mathbb{Q}_\ell$ be the conditional distribution of $\boldsymbol{\xi}$ given $\boldsymbol{\xi}' = \bar{\boldsymbol{\xi}}^\ell$. Then, for every Borel set $A \subseteq \Xi$,

$$\mathbb{Q}(\boldsymbol{\xi} \in A) = \Pi\left((\boldsymbol{\xi}, \boldsymbol{\xi}') \in A \times \Xi\right) = \sum_{\ell=1}^L \Pi\left(\boldsymbol{\xi} \in A \mid \boldsymbol{\xi}' = \bar{\boldsymbol{\xi}}^\ell\right) \widehat{\mathbb{P}}_N(\boldsymbol{\xi}' = \bar{\boldsymbol{\xi}}^\ell) = \sum_{\ell=1}^L \mathbb{Q}_\ell(\boldsymbol{\xi} \in A) \frac{|I_\ell|}{N}.$$

The first equality follows because Π is a joint distribution of $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ with marginals $\mathbb{Q}$ and $\widehat{\mathbb{P}}_N$, respectively. The second equality follows from the law of total probability. The final equality follows from the definitions of $\mathbb{Q}_\ell$ and $\widehat{\mathbb{P}}_N$. Since the above equalities holds for every Borel set, we have shown that

$$\mathbb{Q} = \sum_{\ell=1}^L \frac{|I_\ell|}{N} \mathbb{Q}_\ell.$$

Furthermore, by using similar reasoning as above, we observe that

$$\Pi\left(\|\boldsymbol{\xi} - \boldsymbol{\xi}'\| \le \epsilon_N\right) = \sum_{\ell=1}^L \Pi\left(\|\boldsymbol{\xi} - \boldsymbol{\xi}'\| \le \epsilon_N \mid \boldsymbol{\xi}' = \bar{\boldsymbol{\xi}}^\ell\right) \widehat{\mathbb{P}}_N(\boldsymbol{\xi}' = \bar{\boldsymbol{\xi}}^\ell) = \sum_{\ell=1}^L \mathbb{Q}_\ell(\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}^\ell\| \le \epsilon_N) \frac{|I_\ell|}{N}.$$

Combining the above results, the ambiguity set from (EC.65) can be rewritten as

$$\left\{ \sum_{\ell=1}^{L} \frac{|I_\ell|}{N} \mathbb{Q}_\ell : \begin{array}{l} \sum_{\ell=1}^{L} \mathbb{Q}_\ell(\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}^\ell\| \le \epsilon_N) \frac{|I_\ell|}{N} = 1, \\ \mathbb{Q}_1, \ldots, \mathbb{Q}_L \in \mathcal{P}(\Xi) \end{array} \right\} = \left\{ \sum_{\ell=1}^{L} \frac{|I_\ell|}{N} \mathbb{Q}_\ell : \begin{array}{ll} \mathbb{Q}_\ell(\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}^\ell\| \le \epsilon_N) = 1 & \forall \ell \in [L], \\ \mathbb{Q}_1, \ldots, \mathbb{Q}_L \in \mathcal{P}(\Xi) \end{array} \right\}$$
$$= \left\{ \frac{1}{N} \sum_{j=1}^{N} \mathbb{Q}_j : \begin{array}{ll} \mathbb{Q}_j(\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^j\| \le \epsilon_N) = 1 & \forall j \in [N], \\ \mathbb{Q}_1, \ldots, \mathbb{Q}_N \in \mathcal{P}(\Xi) \end{array} \right\}.$$

The first equality follows because $\mathbb{Q}_\ell(\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}^\ell\| \le \epsilon_N) \le 1$ for each $\ell \in [L]$. The second equality follows because $\mathbb{Q}_\ell(\|\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}^\ell\| \le \epsilon_N) = 1$ if and only if there exists $\mathbb{Q}_j \in \mathcal{P}(\Xi)$ for each $j \in I_\ell$ such that $\mathbb{Q}_j(\|\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}^j\| \le \epsilon_N) = 1$ and $\sum_{j \in I_\ell} \frac{1}{|I_\ell|} \mathbb{Q}_j = \mathbb{Q}_\ell$. This concludes the proof. $\square$

We now present the proof of Proposition 3.

Proposition 3. *Problem* (2) *with uncertainty sets of the form*

$$\mathcal{U}_N^j := \left\{ \boldsymbol{\zeta} \equiv (\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T) \in \Xi : \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}^j\| \le \epsilon_N \right\}$$

is equivalent to ∞*-WDRO.*

Proof. It follows from Lemma EC.5 that the ∞-Wasserstein ambiguity set can be decomposed into separate distributions, each having a support that is contained in $\{\boldsymbol{\zeta} \in \Xi : \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}^j\| \le \epsilon_N\}$ for $j \in [N]$. Of course, these sets are exactly equal to the uncertainty sets from Section 3, and thus Lemma EC.5 implies that the ∞-Wasserstein ambiguity set is equivalent to

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

Therefore, when $\mathcal{A}_N$ is the ∞-Wasserstein ambiguity set and each $\mathcal{U}_N^j$ is a closed balls around $\hat{\boldsymbol{\xi}}^j$ which is intersected with Ξ,

$$\sup_{\mathbb{Q} \in \mathcal{A}_N} \mathbb{E}_{\mathbb{Q}} \left[\sum_{t=1}^{T} \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \right] = \frac{1}{N} \sum_{j=1}^{N} \sup_{\mathbb{Q} \in \mathcal{P}(\mathcal{U}_N^j)} \mathbb{E}_{\mathbb{Q}} \left[\sum_{t=1}^{T} \mathbf{c}_t(\boldsymbol{\xi}) \cdot \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \right]$$
$$= \frac{1}{N} \sum_{j=1}^{N} \sup_{\boldsymbol{\zeta} \in \mathcal{U}_N^j} \sum_{t=1}^{T} \mathbf{c}_t(\boldsymbol{\zeta}) \cdot \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}).$$

Moreover, it similarly follows from Lemma EC.5 that the following inequalities are equivalent:

$$\mathbb{Q}\left(\sum_{t=1}^{T} \mathbf{A}_t(\boldsymbol{\xi}) \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \le \mathbf{b}(\boldsymbol{\xi}) \right) = 1 \quad \forall \mathbb{Q} \in \mathcal{A}_N$$
$$\frac{1}{N} \sum_{j=1}^{N} \mathbb{Q}_j \left(\sum_{t=1}^{T} \mathbf{A}_t(\boldsymbol{\xi}) \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \le \mathbf{b}(\boldsymbol{\xi}) \right) = 1 \quad \forall \mathbb{Q}_j \in \mathcal{P}(\mathcal{U}_N^j),\ j \in [N]$$
$$\mathbb{Q}_j \left(\sum_{t=1}^{T} \mathbf{A}_t(\boldsymbol{\xi}) \mathbf{x}_t(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_{t-1}) \le \mathbf{b}(\boldsymbol{\xi}) \right) = 1 \quad \forall \mathbb{Q}_j \in \mathcal{P}(\mathcal{U}_N^j),\ j \in [N]$$
$$\sum_{t=1}^{T} \mathbf{A}_t(\boldsymbol{\zeta}) \mathbf{x}_t(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_{t-1}) \le \mathbf{b}(\boldsymbol{\zeta}) \qquad \forall \boldsymbol{\zeta} \in \mathcal{U}_N^j,\ j \in [N].$$

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 Proposition 4 from Section 6

In this appendix, we present the proof of Proposition 4.

PROPOSITION 4. *If $p \in [1,\infty)$ and $\epsilon_N > 0$, then a decision rule is feasible for p-WDRO only if*

$$\sum_{t=1}^{T} \mathbf{A}_t(\boldsymbol{\zeta})\mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \le \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \Xi.$$

Proof. Consider any arbitrary $\bar{\boldsymbol{\xi}} \in \Xi$ such that $\bar{\boldsymbol{\xi}} \neq \hat{\boldsymbol{\xi}}^j$ for each $j \in [N]$. Let $\delta_{\bar{\boldsymbol{\xi}}}$ denote the Dirac delta distribution which satisfies $\delta_{\bar{\boldsymbol{\xi}}}(\boldsymbol{\xi} = \bar{\boldsymbol{\xi}}) = 1$, and let $\widehat{\mathbb{P}}_N := \frac{1}{N}\sum_{j=1}^{N} \delta_{\hat{\boldsymbol{\xi}}^j}$ 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

$$\mathbb{Q}_{\bar{\boldsymbol{\xi}}}^{\lambda} := (1-\lambda)\widehat{\mathbb{P}}_N + \lambda \delta_{\bar{\boldsymbol{\xi}}}.$$

We recall the definition of the p-Wasserstein distance between $\widehat{\mathbb{P}}_N$ and $\mathbb{Q}_{\bar{\boldsymbol{\xi}}}^{\lambda}$:

$$\mathsf{d}_p\left(\widehat{\mathbb{P}}_N, \mathbb{Q}_{\bar{\boldsymbol{\xi}}}^{\lambda}\right) = \inf\left\{ \left(\int_{\Xi\times\Xi} \|\boldsymbol{\xi}-\boldsymbol{\xi}'\|^p \, d\Pi(\boldsymbol{\xi},\boldsymbol{\xi}')\right)^{\frac{1}{p}} : \begin{array}{l} \Pi \text{ is a joint distribution of } \boldsymbol{\xi} \text{ and } \boldsymbol{\xi}' \\ \text{with marginals } \widehat{\mathbb{P}}_N \text{ and } \mathbb{Q}_{\bar{\boldsymbol{\xi}}}^{\lambda}, \text{ respectively} \end{array} \right\}. \tag{EC.66}$$

Consider a feasible joint distribution $\bar{\Pi}$ for the above optimization problem in which $\boldsymbol{\xi}' \sim \mathbb{Q}_{\bar{\boldsymbol{\xi}}}^{\lambda}$, $\boldsymbol{\xi}'' \sim \widehat{\mathbb{P}}_N$, and

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

Indeed, we readily verify that the marginal distributions of $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$ are $\widehat{\mathbb{P}}_N$ and $\mathbb{Q}_{\bar{\boldsymbol{\xi}}}^{\lambda}$, respectively, and thus this joint distribution is feasible for the optimization problem in (EC.66). Moreover,

$$\begin{aligned}
\mathsf{d}_p\left(\widehat{\mathbb{P}}_N, \mathbb{Q}_{\bar{\boldsymbol{\xi}}}^{\lambda}\right) &\le \left(\int_{\Xi\times\Xi} \|\boldsymbol{\xi}-\boldsymbol{\xi}'\|^p \, d\bar{\Pi}(\boldsymbol{\xi},\boldsymbol{\xi}')\right)^{\frac{1}{p}} \\
&= \left(\int_{\Xi\times\Xi} \|\boldsymbol{\xi}-\boldsymbol{\xi}'\|^p \, \mathbb{I}\left\{\boldsymbol{\xi}' = \bar{\boldsymbol{\xi}}\right\} d\bar{\Pi}(\boldsymbol{\xi},\boldsymbol{\xi}') + \underbrace{\int_{\Xi\times\Xi} \|\boldsymbol{\xi}-\boldsymbol{\xi}'\|^p \, \mathbb{I}\left\{\boldsymbol{\xi}' \neq \bar{\boldsymbol{\xi}}\right\} d\bar{\Pi}(\boldsymbol{\xi},\boldsymbol{\xi}')}_{=0}\right)^{\frac{1}{p}} \\
&= \left(\frac{1}{N}\sum_{j=1}^{N} \lambda \left\|\hat{\boldsymbol{\xi}}^j - \bar{\boldsymbol{\xi}}\right\|^p\right)^{\frac{1}{p}}.
\end{aligned}$$

The inequality follows since $\bar{\Pi}$ is a feasible but possibly suboptimal joint distribution for the optimization problem in (EC.66). The first equality follows from splitting the integral into two cases, and observing that the second case equals zero since $\boldsymbol{\xi} = \boldsymbol{\xi}'$ whenever $\boldsymbol{\xi}' \neq \bar{\boldsymbol{\xi}}$. The final equality follows because $\boldsymbol{\xi} = \boldsymbol{\xi}''$ whenever $\boldsymbol{\xi}' = \bar{\boldsymbol{\xi}}$, and $\boldsymbol{\xi}''$ is distributed uniformly over the historical sample paths. Thus, for any arbitrary choice of $\bar{\boldsymbol{\xi}} \in \Xi$, we have shown that $\mathbb{Q}_{\bar{\boldsymbol{\xi}}}^{\lambda}$ is contained in the p-Wasserstein ambiguity set whenever $\lambda \in (0,1)$ satisfies

$$\begin{aligned}
\left(\frac{1}{N}\sum_{j=1}^{N} \lambda \left\|\hat{\boldsymbol{\xi}}^j - \bar{\boldsymbol{\xi}}\right\|^p\right)^{\frac{1}{p}} &\le \epsilon_N \\
\frac{1}{N}\sum_{j=1}^{N} \lambda \left\|\hat{\boldsymbol{\xi}}^j - \bar{\boldsymbol{\xi}}\right\|^p &\le \epsilon_N^p \\
\lambda &\le \frac{\epsilon_N^p}{\frac{1}{N}\sum_{j=1}^{N} \left\|\hat{\boldsymbol{\xi}}^j - \bar{\boldsymbol{\xi}}\right\|^p}.
\end{aligned}$$

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

$$\sum_{t=1}^{T} \mathbf{A}_t(\boldsymbol{\xi})\mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1}) \le \mathbf{b}(\boldsymbol{\xi}) \quad \mathbb{Q}\text{-a.s.}, \ \forall \mathbb{Q} \in \mathcal{A}_N. \tag{EC.67}$$

Let $\mathcal{A}_N$ be the p-Wasserstein ambiguity set for $1 \le p < \infty$ and $\epsilon_N > 0$. Then, for any arbitrary $\bar{\boldsymbol{\xi}} \in \Xi$, there exists a $\lambda \in (0,1)$ such that $\mathbb{Q}^{\lambda}_{\bar{\boldsymbol{\xi}}}$ is contained in $\mathcal{A}_N$, and so it follows from (EC.67) that the decision rule must satisfy

$$\sum_{t=1}^{T} \mathbf{A}_t(\bar{\boldsymbol{\xi}})\mathbf{x}_t(\bar{\boldsymbol{\xi}}_1,\ldots,\bar{\boldsymbol{\xi}}_{t-1}) \le \mathbf{b}(\bar{\boldsymbol{\xi}}).$$

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

$$\sum_{t=1}^{T} \mathbf{A}_t(\boldsymbol{\zeta})\mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \le \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \Xi,$$

which is what we wished to show. □

Appendix I: Linear Decision Rules for Problem (6) with 1-Wasserstein Ambiguity Sets

In this appendix, we present a reformulation of linear decision rules for Problem (6) using the 1-Wasserstein ambiguity set. The performance of this data-driven approach is illustrated in Section 8.

We first review the necessary notation. Following Section 5, we focus on a specific case of Problem (6) of the form

$$\begin{aligned} &\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} && \sup_{\mathbb{Q}\in\mathcal{A}_N} \mathbb{E}_{\mathbb{Q}}\left[\sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})\right] \\ &\text{subject to} && \sum_{t=1}^{T} \mathbf{A}_t \mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1}) \le \mathbf{b}(\boldsymbol{\xi}) \quad \mathbb{Q}\text{-a.s.}, \ \forall \mathbb{Q}\in\mathcal{A}_N, \end{aligned} \tag{EC.68}$$

in which $\mathbf{A}_t(\boldsymbol{\xi})$ and $\mathbf{c}_t(\boldsymbol{\xi})$ do not depend on the stochastic process. The ambiguity set is constructed as

$$\mathcal{A}_N = \left\{ \mathbb{Q} \in \mathcal{P}(\Xi) : \, \mathsf{d}_1(\mathbb{Q}, \widehat{\mathbb{P}}_N) \le \epsilon_N \right\},$$

where $\widehat{\mathbb{P}}_N$ is the empirical distribution of the historical data, $\epsilon_N \ge 0$ is the robustness parameter, and the 1-Wasserstein distance between two distributions is given by

$$\mathsf{d}_1(\mathbb{Q},\mathbb{Q}') = \inf \left\{ \int_{\Xi\times\Xi} \|\boldsymbol{\xi} - \boldsymbol{\xi}'\| \, d\Pi(\boldsymbol{\xi},\boldsymbol{\xi}') : \begin{array}{l} \Pi \text{ is a joint distribution of } \boldsymbol{\xi} \text{ and } \boldsymbol{\xi}' \\ \text{with marginals } \mathbb{Q} \text{ and } \mathbb{Q}', \text{ respectively} \end{array} \right\}.$$

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 Proposition 4 in Section 6 that Problem (EC.68) is equivalent to

$$\begin{aligned} &\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} && \sup_{\mathbb{Q}\in\mathcal{A}_N} \mathbb{E}_{\mathbb{Q}}\left[\sum_{t=1}^{T} \mathbf{c}_t \cdot \mathbf{x}_t(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_{t-1})\right] \\ &\text{subject to} && \sum_{t=1}^{T} \mathbf{A}_t \mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) \le \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta}\in\Xi. \end{aligned} \tag{EC.69}$$

We next present an extension of the linear decision rule approach to Problem (EC.69), in which we restrict the space of decision rules to those of the form

$$\mathbf{x}_t(\boldsymbol{\zeta}_1,\ldots,\boldsymbol{\zeta}_{t-1}) = \mathbf{x}_{t,0} + \sum_{s=1}^{t-1} \mathbf{X}_{t,s}\boldsymbol{\zeta}_s.$$

The resulting approximation of Problem (EC.69) is given by

$$\begin{aligned} &\text{minimize} && \sup_{\mathbb{Q}\in\mathcal{A}_N} \mathbb{E}_{\mathbb{Q}}\left[\sum_{t=1}^{T} \mathbf{c}_t\cdot\left(\mathbf{x}_{t,0} + \sum_{s=1}^{t-1}\mathbf{X}_{t,s}\boldsymbol{\xi}_s\right)\right] \\ &\text{subject to} && \sum_{t=1}^{T}\mathbf{A}_t\left(\mathbf{x}_{t,0} + \sum_{s=1}^{t-1}\mathbf{X}_{t,s}\boldsymbol{\zeta}_s\right) \le \mathbf{b}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta}\in\Xi, \end{aligned} \tag{EC.70}$$

where the decision variables are $\mathbf{x}_{t,0}\in\mathbb{R}^{n_t}$ and $\mathbf{X}_{t,s}\in\mathbb{R}^{n_t\times d_s}$ for all $1\le s<t$.

In the remainder of this appendix, we develop a tractable reformulation of Problem (EC.70). Our reformulation, which will use similar duality techniques to those presented in Section 5, is presented as Theorem EC.3. Our reformulation requires the following assumption:

ASSUMPTION EC.1. *The set $\Xi\subseteq\mathbb{R}^d$ is a nonempty multi-dimensional box of the form $[\boldsymbol{\ell},\mathbf{u}]$, where any component of $\boldsymbol{\ell}$ might be $-\infty$ and any component of $\mathbf{u}$ may be ∞. Moreover, the norm in the 1-Wasserstein distance is equal to $\|\cdot\|_1$.*

We now present the reformulation of Problem (EC.70) given Assumption EC.1.

THEOREM EC.3. *If Assumption EC.1 holds, then Problem* (EC.70) *can be reformulated by adding at most $O(md)$ additional continuous decision variables and $O(md)$ additional linear constraints. The reformulation is given by*

$$\begin{aligned} &\text{minimize} && \lambda\epsilon_N + \frac{1}{N}\sum_{j=1}^{N}\left(\sum_{t=1}^{T}\mathbf{c}_t\cdot\left(\mathbf{x}_{t,0}+\sum_{s=1}^{t-1}\mathbf{X}_{t,s}\hat{\boldsymbol{\xi}}_s^j\right) + \boldsymbol{\alpha}\cdot(\mathbf{u}-\hat{\boldsymbol{\xi}}^j) + \boldsymbol{\beta}(\hat{\boldsymbol{\xi}}^j-\boldsymbol{\ell})\right) \\ &\text{subject to} && \left\|\sum_{s=t+1}^{T}(\mathbf{X}_{s,t})^\intercal\mathbf{c}_s - \boldsymbol{\alpha}_t + \boldsymbol{\beta}_t\right\|_\infty \le \lambda && t\in[T] \\ &&& \mathbf{M}_t - \boldsymbol{\Lambda}_t = -\mathbf{B}_t + \sum_{s=t+1}^{T}\mathbf{A}_s\mathbf{X}_{s,t} && t\in[T] \\ &&& \sum_{t=1}^{T}(\mathbf{M}_t\mathbf{u}_t - \boldsymbol{\Lambda}_t\boldsymbol{\ell}_t + \mathbf{A}_t\mathbf{x}_{t,0}) \le \mathbf{b}^0, \end{aligned}$$

where the auxilary decision variables are $\boldsymbol{\alpha} \coloneqq (\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_T), \boldsymbol{\beta} \coloneqq (\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_T)\in\mathbb{R}^d_+$, as well as $\mathbf{M} \coloneqq (\mathbf{M}_1,\ldots,\mathbf{M}_T), \boldsymbol{\Lambda} \coloneqq (\boldsymbol{\Lambda}_1,\ldots,\boldsymbol{\Lambda}_T)\in\mathbb{R}^{m\times d}_+$.

Proof. Following similar reasoning to Theorem 4, the constraints

$$\sum_{t=1}^{T}\mathbf{A}_t\left(\mathbf{x}_{t,0}+\sum_{s=1}^{t-1}\mathbf{X}_{t,s}\boldsymbol{\zeta}_s\right) \le \mathbf{b}^0 + \sum_{t=1}^{T}\mathbf{B}_t\boldsymbol{\zeta}_t \quad \forall\boldsymbol{\zeta}\in\Xi$$

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

$$\begin{aligned} \sum_{t=1}^{T} (\mathbf{M}_t \mathbf{u}_t - \boldsymbol{\Lambda}_t \boldsymbol{\ell}_t + \mathbf{A}_t \mathbf{x}_{t,0}) \leq \mathbf{b}^0, \\ \mathbf{M}_t - \boldsymbol{\Lambda}_t = \sum_{s=t+1}^{T} \mathbf{A}_s \mathbf{X}_{s,t} - \mathbf{B}_t, \qquad t \in [T]. \end{aligned}$$

The remainder of the proof focuses on the objective function. Note that for any fixed solution to Problem (EC.70) one can define a function $f : \Xi \to \mathbb{R}$ as follows

$$f(\boldsymbol{\zeta}) = \sum_{t=1}^{T} \mathbf{c}_t \cdot \left(\mathbf{x}_{t,0} + \sum_{s=1}^{t-1} \mathbf{X}_{t,s} \boldsymbol{\zeta}_s \right).$$

It follows from Assumption EC.1 that $\Xi \subseteq \mathbb{R}^d$ is nonempty, convex, and closed, and $-f(\cdot)$ is proper, convex, and lower semicontinuous on Ξ, thus satisfying Mohajerin Esfahani and Kuhn (2018, Assumption 4.1). Therefore, we conclude from Mohajerin Esfahani and Kuhn (2018, Equation 12b) that

$$\begin{aligned} \sup_{\mathbb{Q} \in \mathcal{A}_N} \mathbb{E}_{\mathbb{Q}} \left[\sum_{t=1}^{T} \mathbf{c}_t \cdot \left(\mathbf{x}_{t,0} + \sum_{s=1}^{t-1} \mathbf{X}_{t,s} \boldsymbol{\xi}_s \right) \right] &= \sup_{\mathbb{Q} \in \mathcal{A}_N} \mathbb{E}_{\mathbb{Q}} \left[f(\boldsymbol{\xi}) \right] \\ &= \inf_{\lambda \geq 0} \lambda \epsilon_N + \frac{1}{N} \sum_{j=1}^{N} \sup_{\boldsymbol{\zeta} \in \Xi} \left\{ f(\boldsymbol{\zeta}) - \lambda \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}^j\|_1 \right\} \\ &= \inf_{\lambda \geq 0} \lambda \epsilon_N + \frac{1}{N} \sum_{j=1}^{N} \underbrace{\sup_{\boldsymbol{\zeta} \in \Xi} \left\{ \sum_{t=1}^{T} \mathbf{c}_t \cdot \left(\mathbf{x}_{t,0} + \sum_{s=1}^{t-1} \mathbf{X}_{t,s} \boldsymbol{\zeta}_s \right) - \lambda \|\boldsymbol{\zeta} - \hat{\boldsymbol{\xi}}^j\|_1 \right\}}_{\gamma_j}. \end{aligned} \tag{EC.71}$$

We now reformulate the expression γ_j for each $j \in [N]$. Indeed, it follows from strong duality for linear programming that

$$\begin{aligned} \gamma_j = \quad \underset{\boldsymbol{\alpha},\, \boldsymbol{\beta} \in \mathbb{R}_+^d}{\text{minimize}} \quad & \sum_{t=1}^{T} \Big(\mathbf{c}_t \cdot \mathbf{x}_{t,0} + \boldsymbol{\alpha}_t \cdot (\mathbf{u}_t - \hat{\boldsymbol{\xi}}_t^j) + \boldsymbol{\beta}_t \cdot (\hat{\boldsymbol{\xi}}_t^j - \boldsymbol{\ell}_t) \Big) \\ \text{subject to} \quad & \left\| \sum_{s=t+1}^{T} (\mathbf{X}_{s,t})^\intercal \mathbf{c}_s - \boldsymbol{\alpha}_t + \boldsymbol{\beta}_t \right\|_\infty \leq \lambda, \quad t \in [T]. \end{aligned} \tag{EC.72}$$

Remark: For any index l such that $u_l = \infty$ (alternatively, $\ell_l = -\infty$), the corresponding decision variable α_l (alternatively, β_l) should be set to zero and the term $\alpha_l(u_l - \hat{\xi}_l^j)$ (alternatively, $\beta_l(\hat{\xi}_l^j - \ell_l)$) should be dropped from the objective.

Note that problem (EC.72) is component-wise separable to d problems of the form

$$\begin{aligned} \underset{\alpha_l,\, \beta_l \in \mathbb{R}_+}{\text{minimize}} \quad & \alpha_l(u_l^k - \hat{\xi}_l^j) + \beta_l(\hat{\xi}_l^j - \ell_l) \\ \text{subject to} \quad & |g_l - \alpha_l + \beta_l| \leq \lambda, \end{aligned} \tag{EC.73}$$

where $\mathbf{g} := \left(\sum_{s=2}^{T} (\mathbf{X}_{s,1})^\intercal \mathbf{c}_s, \sum_{s=3}^{T} (\mathbf{X}_{s,2})^\intercal \mathbf{c}_s, \ldots, (\mathbf{X}_{T,T-1})^\intercal \mathbf{c}_T, 0 \right) \in \mathbb{R}^d$. Moreover, $\hat{\boldsymbol{\xi}}^j \in \Xi$ implies that both $(u_l - \hat{\xi}_l^j)$ and $(\hat{\xi}_l^j - \ell_l)$ are nonnegative, and so for any fixed λ and g_l, an optimal solution of (EC.73) is given by $\alpha_l = \max\{g_l - \lambda, 0\}$ and $\beta_l = \max\{-g_l - \lambda, 0\}$ (their corresponding minimal values). This solution is independent of the value of $\hat{\xi}_l^j$, and therefore, the same variables $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be used in (EC.72) for all values of $j \in [N]$. Combining (EC.71) and (EC.72) and plugging the result to the objective function of (EC.70), we obtain the desired formulation. $\square$

Appendix J: Supplement to Section 7

J.1. Reformulation of the Three-Stage Inventory Replenishment Problem

Let $I_{2r}^{-}(\boldsymbol{\xi}_1) = \max\{0, \xi_{1r} - Q_{1r}\}$ denote the demand at retailer r from the first half of the week that exceeded the initial inventory of the retailer. With these auxiliary decision rules, we observe that the three-stage stochastic nonlinear optimization problem from Section 7.1 is equivalent to

$$\begin{aligned}
\underset{\mathbf{Q}\ge\mathbf{0},\mathbf{z}\in\{0,1\}^R,\mathbf{I}}{\text{minimize}} \quad & \mathbb{E}\Bigg[c\left(Q_{10} + \sum_{r=1}^R Q_{1r}\right) + h\left(Q_{10} - \sum_{r=1}^R Q_{2r}(\boldsymbol{\xi}_1)\right) + b\left(\sum_{r=1}^R I_{2r}^{-}(\boldsymbol{\xi}_1)\right) \\
& \quad + f\left(\sum_{r=1}^R z_r(\boldsymbol{\xi}_1)\right) + b\left(\sum_{r=1}^R \max\{0, -I_{3r}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)\}\right) + h\left(\sum_{r=1}^R \max\{0, I_{3r}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)\}\right)\Bigg] \\
\text{subject to} \quad & \sum_{r=1}^R Q_{2r}(\boldsymbol{\xi}_1) \le Q_{10} && \text{a.s.} \\
& I_{2r}^{-}(\boldsymbol{\xi}_1) = \max\{0, \xi_{1r} - Q_{1r}\} && \forall r \in [R], \text{ a.s.} \\
& I_{3r}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) = Q_{1r} - \xi_{1r} + I_{2r}^{-}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) + Q_{2r}(\boldsymbol{\xi}_1) - \xi_{2r} && \forall r \in [R], \text{ a.s.} \\
& z_r(\boldsymbol{\xi}_1)\mathcal{M} \ge Q_{2r}(\boldsymbol{\xi}_1) && \forall r \in [R], \text{ a.s.}
\end{aligned}$$

After substituting the equality of $I_{3r}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)$ into the objective function, and after adding epigraph decision rules, we obtain the following simplified formulation:

$$\begin{aligned}
\underset{\mathbf{Q}\ge\mathbf{0},\mathbf{z}\in\{0,1\}^R,\mathbf{I}^-,\mathbf{v}}{\text{minimize}} \quad & \mathbb{E}\left[c\left(Q_{10} + \sum_{r=1}^R Q_{1r}\right) + hQ_{10} + \sum_{r=1}^R v_r(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) + f\sum_{r=1}^R z_r(\boldsymbol{\xi}_1)\right] \\
\text{subject to} \quad & \sum_{r=1}^R Q_{2r}(\boldsymbol{\xi}_1) \le Q_{10} && \text{a.s.} \\
& I_{2r}^{-}(\boldsymbol{\xi}_1) = \max\{0, \xi_{1r} - Q_{1r}\} && \forall r \in [R], \text{ a.s.} \\
& v_r(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) \ge b(\xi_{2r} + \xi_{1r} - Q_{2r}(\boldsymbol{\xi}_1) - Q_{1r}) - hQ_{2r}(\boldsymbol{\xi}_1) && \forall r \in [R], \text{ a.s.} \\
& v_r(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) \ge (h+b)I_{2r}^{-}(\boldsymbol{\xi}_1) + h(Q_{1r} - \xi_{1r} - \xi_{2r}) && \forall r \in [R], \text{ a.s.} \\
& z_r(\boldsymbol{\xi}_1)\mathcal{M} \ge Q_{2r}(\boldsymbol{\xi}_1) && \forall r \in [R], \text{ a.s.}
\end{aligned}$$

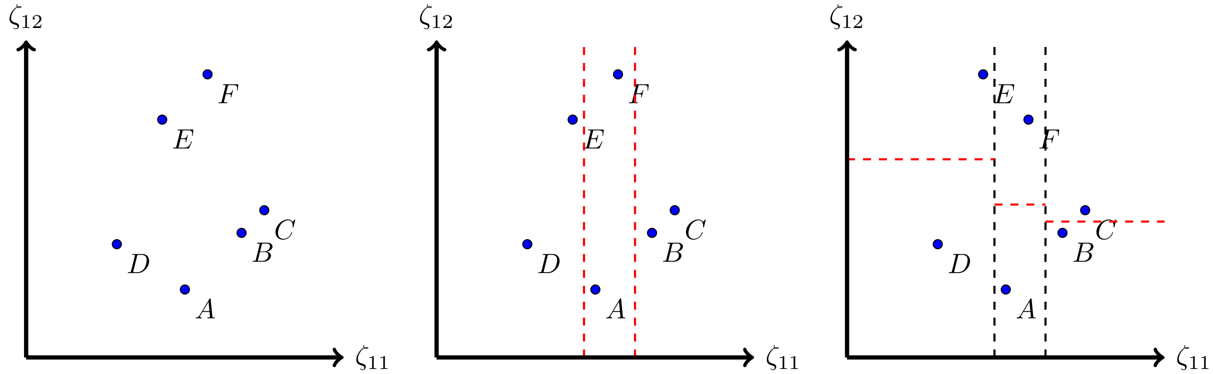
Plugging in the equality of the auxiliary decision rules $I_{2r}^{-}(\boldsymbol{\xi}_1)$ into the remaining constraints, and eliminating the maximization by splitting the relevant constraint, we conclude our derivation of Problem (7).

J.2. Heuristic partitioning algorithm

In this appendix, we describe the heuristic partitioning algorithm for finite adaptability which is used in our numerical experiments in Section 7. The goal of the algorithm is to construct a partition comprised of hyper-rectangular regions of the form $P^1 := [\boldsymbol{\ell}^1, \mathbf{u}^1] \times \mathbb{R}_+^R, \ldots, P^N := [\boldsymbol{\ell}^N, \mathbf{u}^N] \times \mathbb{R}_+^R$ such that the inclusion $\hat{\boldsymbol{\xi}}^j \in P^j$ is satisfied for each sample path $j \in [N]$. The output of the partitioning algorithm from this section are thus the vectors $\boldsymbol{\ell}^1, \mathbf{u}^1, \ldots, \boldsymbol{\ell}^N, \mathbf{u}^N \in [0,\infty]^R$ which define the partition of the set $\Xi := \mathbb{R}_+^{2R}$.

Our algorithm is comprised of the following steps, which are formalized in Algorithm 1 and visualized in Figure EC.1. We first define $M := \lceil N^{\frac{1}{R}} \rceil$ as the smallest integer which satisfies the inequality $M^R \ge N$.

Figure EC.1 Illustration of the partitioning algorithm



Note. **Left**: This shows the $N=6$ historical sample paths in an example with $R=2$ retailers. **Center**: In the first iteration ($r=1$) of the heuristic partitioning algorithm, there are $N=6$ historical sample paths. Since there are $R=2$ retailers, we subdivide the historical sample paths into $M=3$ regions along the demand of the first retailer. **Right**: In the second iteration ($r=2$), we subdivide each of the 3 regions along the demand of the second retailer. The regions comprise the partition which is returned by the algorithm.

We then iterate over the indices $r=1,2,\ldots,R$. In each iteration r, we start with a partition of Ξ and then subdivide each region in that partition into at most M smaller regions by adding cuts along the first-stage demand ζ_{1r}. The borders of the new regions are determined by the historical data points such that each region contains approximately the same number of data points and the region borders are the furthest possible from the closest data point in the ζ_{1r} dimension. The progression of the algorithm is illustrated in Figure EC.1.

When the algorithm concludes its Rth iteration, we have obtained a partition $\mathcal{P}=\{P^1,\ldots,P^N\}$ of Ξ with exactly N regions such that each $\hat{\boldsymbol{\xi}}_1^j$ lies in its own region. Moreover, since Ξ is a hyperrectangle, each region P^j is also a hyperrectangle. This concludes the description of our heuristic partitioning algorithm. We note that, for simplicity, the version of the algorithm presented here ignores cases where there exist $j_1 \neq j_2 \in [N]$ and r such that $\hat{\boldsymbol{\xi}}_{1r}^{j_1} = \hat{\boldsymbol{\xi}}_{1r}^{j_2}$, which may happen if the distribution is not continuous; however, this is not the case in our numerical experiments, and, furthermore, our algorithm can be easily adjusted to address such ties.

J.3. Reformulation of the robust optimization problem

In this appendix, we present the derivation of Problem (9) from Section 7.2. Indeed, we recall from Section 7.2 that the uncertainty sets $\mathcal{U}_N^1,\ldots,\mathcal{U}_N^N$ are hyperrectangles, and it follows from Appendix J.2 that the regions $P^1,\ldots,P^K \subseteq \Xi$ that are obtained from our heuristic partitioning algorithm are hyperrectangles as well.

Algorithm 1 Partitioning algorithm

Input: Ξ, N, $\{\hat{\boldsymbol{\xi}}^1, \dots, \hat{\boldsymbol{\xi}}^N\}$

Output: Partition $\mathcal{P}$

Initialize $M = \lceil N^{\frac{1}{R}} \rceil$, $\mathcal{P} = \{\Xi\}$

for $r := 1, \dots, R$ **do**

  for all $P \in \mathcal{P}$ **do**

    Find $J = \{j : \hat{\boldsymbol{\xi}}^j \in P\}$

    Let $\{j_{(k)}\}_{k=1,\dots,|J|}$ be an ordering of the indexes in J such that

$$\hat{\xi}_{1r}^{j_{(k)}} \le \hat{\xi}_{1r}^{j_{(k+1)}}, k \in [|J| - 1].$$

    Update $\mathcal{P} = \mathcal{P} \setminus \{P\}$.

    Set $K = \min\{M, |J|\}$ and $k_0 = 0$

    for all $l := 1, \dots, K$ **do**

      Set $k_l = \max\{\lceil |J| l / K \rceil, k_{l-1} + 1\}$

      if $k < |J|$ **then**

        $\bar{\zeta}^l = \frac{\hat{\xi}_{1r}^{j_{(k)}} + \hat{\xi}_{1r}^{j_{(k+1)}}}{2}$

        Set $P^l = \{\boldsymbol{\zeta} \in P : \zeta_{1r} \le \bar{\zeta}^l\}$ and update $P = P \setminus P^l$.

      else

        Set $K = l$ and $P^K = P$

      end if

    end for

    Update $\mathcal{P} = \mathcal{P} \cup \{P^1, \dots, P^K\}$.

  end for

end for

Consequently, the approximation of the robust optimization problem using finite adaptability is given by

$$
\begin{aligned}
\underset{\substack{\mathbf{v},\,\mathbf{Q}_1\geq\mathbf{0},\\ \mathbf{Q}_2^k\geq 0,\mathbf{z}^k\in\{0,1\}^R}}{\text{minimize}} \quad & c\left(Q_{10}+\sum_{r=1}^R Q_{1r}\right)+hQ_{10}+\frac{1}{N}\sum_{j=1}^N \max_{k\in\mathcal{K}_j,\boldsymbol{\zeta}\in\mathcal{U}_N^j\cap P^k}\left\{\sum_{r=1}^R v_r(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2)+f\sum_{r=1}^R z_r^k\right\}\\
\text{subject to} \quad & \sum_{r=1}^R Q_{2r}^k \leq Q_{10} && \forall k\in[K]\\
& v_r(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2)\geq b(\zeta_{2r}+\zeta_{1r}-Q_{2r}^k-Q_{1r})-hQ_{2r}^k && \forall r\in[R],\ k\in[K],\ \boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j\cap P^k\\
& v_r(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2)\geq h(Q_{1r}-\zeta_{1r}-\zeta_{2r}) && \forall r\in[R],\ k\in[K],\ \boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j\cap P^k\\
& v_r(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2)\geq b\,(\zeta_{1r}-Q_{1r})-h\zeta_{2r} && \forall r\in[R],\ k\in[K],\ \boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j\cap P^k\\
& z_r^k\mathcal{M}\geq Q_{2r}^k && \forall r\in[R],\ k\in[K],
\end{aligned}
\tag{EC.74}
$$

where $\mathcal{K}_j := \{k\in[N] : \mathcal{U}_N^j\cap P^k\neq\emptyset\}$ contains the indices of regions $P^1,\ldots,P^K$ that intersect the uncertainty set $\mathcal{U}_N^j$. In the remainder of this appendix, we show that the above optimization problem is equivalent to Problem (9).

We first observe that there is an optimal solution to Problem (EC.74) in which each decision rule $v_r(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2)$ is equivalent to a function that depends only on ζ_{1r} and ζ_{2r}. Therefore, since each $\mathcal{U}_N^j\cap P^k$ is a hyperrectangle, we observe that

$$
\max_{k\in\mathcal{K}_j,\boldsymbol{\zeta}\in\mathcal{U}_N^j\cap P^k}\left\{\sum_{r=1}^R v_r(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2)+f\sum_{r=1}^R z_r^k\right\}=\max_{k\in\mathcal{K}_j}\left\{\sum_{r=1}^R\left(\max_{\boldsymbol{\zeta}\in\mathcal{U}_N^j\cap P^k} v_r(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2)+fz_r^k\right)\right\}.
$$

It follows from the above observation that Problem (EC.74) is equivalent to

$$
\begin{aligned}
\underset{\substack{\mathbf{v},\,\mathbf{Q}_1\geq\mathbf{0},\\ \mathbf{Q}_2^k\geq 0,\mathbf{z}^k\in\{0,1\}^R\\ \mathbf{u}^{j,k},\mathbf{v}^j\in\mathbb{R}^R}}{\text{minimize}} \quad & c\left(Q_{10}+\sum_{r=1}^R Q_{1r}\right)+hQ_{10}+\frac{1}{N}\sum_{j=1}^N v_r^j\\
\text{subject to} \quad & \sum_{r=1}^R Q_{2r}^k\leq Q_{10} && \forall k\in[K]\\
& v_r^j\geq\sum_{r=1}^R\left(u_r^{j,k}+fz_r^k\right) && \forall r\in[R],\ j\in[N],\ k\in\mathcal{K}_j\\
& u_r^{j,k}\geq b(\zeta_{2r}+\zeta_{1r}-Q_{2r}^k-Q_{1r})-hQ_{2r}^k && \forall r\in[R],\ k\in[K],\ \boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j\cap P^k\\
& u_r^{j,k}\geq h(Q_{1r}-\zeta_{1r}-\zeta_{2r}) && \forall r\in[R],\ k\in[K],\ \boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j\cap P^k\\
& u_r^{j,k}\geq b\,(\zeta_{1r}-Q_{1r})-h\zeta_{2r} && \forall r\in[R],\ k\in[K],\ \boldsymbol{\zeta}\in\cup_{j=1}^N\mathcal{U}_N^j\cap P^k\\
& z_r^k\mathcal{M}\geq Q_{2r}^k && \forall r\in[R],\ k\in[K],
\end{aligned}
$$

where at optimality each auxiliary decision variable $u_r^{j,k}$ will satisfy the equality $u_r^{j,k}=\max_{\boldsymbol{\zeta}\in\mathcal{U}^j\cap P^k} v_r(\boldsymbol{\zeta}_1,\boldsymbol{\zeta}_2)$. Applying the standard 'robust counterpart' reformulation technique, we observe that the above optimization problem can be rewritten as

$$
\begin{array}{lll}
\underset{\substack{\mathbf{v}^j,\, \mathbf{Q}_1 \geq \mathbf{0}, \\ \mathbf{Q}_2^k \geq 0, \mathbf{z}^k \in \{0,1\}^R \\ \mathbf{u}^{j,k}, \mathbf{v}^j \in \mathbb{R}^R}}{\text{minimize}} & c\left(Q_{10} + \displaystyle\sum_{r=1}^R Q_{1r}\right) + hQ_{10} + \dfrac{1}{N}\displaystyle\sum_{j=1}^N \sum_{r=1}^R v_r^j & \\
\text{subject to} & \displaystyle\sum_{r=1}^R Q_{2r}^k \leq Q_{10} & \forall k \in [K] \\
& v_r^j \geq \displaystyle\sum_{r=1}^R \left(u_r^{j,k} + f z_r^k\right) & \forall r \in [R],\, j \in [N],\, k \in \mathcal{K}_j \\
& u_r^{j,k} \geq \displaystyle\max_{\boldsymbol{\zeta} \in \mathcal{U}_N^j \cap P^k} b(\zeta_{2r} + \zeta_{1r} - Q_{2r}^k - Q_{1r}) - hQ_{2r}^k & \forall r \in [R],\, j \in [N],\, k \in \mathcal{K}_j \\
& u_r^{j,k} \geq \displaystyle\max_{\boldsymbol{\zeta} \in \mathcal{U}_N^j \cap P^k} h(Q_{1r} - \zeta_{1r} - \zeta_{2r}) & \forall r \in [R],\, j \in [N],\, k \in \mathcal{K}_j \\
& u_r^{j,k} \geq \displaystyle\max_{\boldsymbol{\zeta} \in \mathcal{U}_N^j \cap P^k} b\left(\zeta_{1r} - Q_{1r}\right) - h\zeta_{2r} & \forall r \in [R],\, j \in [N],\, k \in \mathcal{K}_j \\
& z_r^k \mathcal{M} \geq Q_{2r}^k & \forall r \in [R],\, k \in [K].
\end{array}
$$

Let us define a lower bound $\underline{\zeta}_{tr}^{jk} := \min_{\boldsymbol{\zeta} \in \mathcal{U}_N^j \cap P^k} \zeta_{tr}$ and an upper bound $\bar{\zeta}_{tr}^{jk} := \max_{\boldsymbol{\zeta} \in \mathcal{U}_N^j \cap P^k} \zeta_{tr}$ for each period $t \in \{1,2\}$, retailer $r \in [R]$, sample path $j \in [N]$, and region $k \in \mathcal{K}_j$. Since each set $\mathcal{U}_N^j \cap P^k$ is a hyperrectangle, and since the holding costs and backlogging costs are nonnegative ($h, b \geq 0$), we see that the above optimization problem is equivalent to Problem (9).

J.4. Supplement to Section 7.3

We conclude the present Appendix J with additional numerical results which were omitted from Section 7. In Figure EC.2, we show the impact of having a fixed cost of $f = 0.1$ versus not having a fixed cost on the replenishment decision rules obtained from SRO-FA. In Figures EC.3 and EC.4, we show the impact of the robustness parameter on SRO-FA for various numbers of retailers and sizes of training datasets with and without fixed costs.

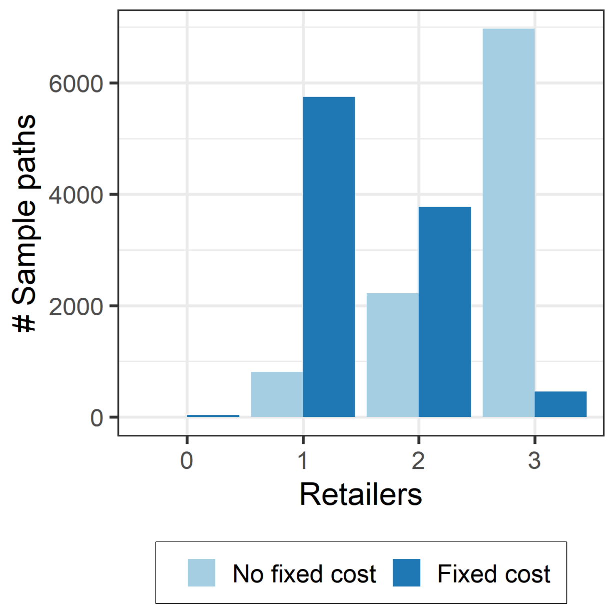
Appendix K: Supplement to Section 8.3

In this appendix, we provide supplemental numerical results for the multi-stage stochastic inventory management problem from Section 8. Specifically, the aim of this appendix is to evaluate the impact of the projection procedure, described at the end of Section 8.2, on the out-of-sample costs of SRO-LDR and SAA-LDR reported in Table 1.

Following the same notation in Section 8.2, let $\mathbf{x}^{\mathcal{A},i,\ell} = (x_1^{\mathcal{A},i,\ell}, \ldots, x_T^{\mathcal{A},i,\ell})$ be the production quantities obtained when the decision rule from approach $\mathcal{A}$ on training dataset ℓ is applied to the ith sample path in the testing dataset. For each approach $\mathcal{A}$ and training dataset ℓ, the probability that the resulting decision rule is *feasible* is approximated by

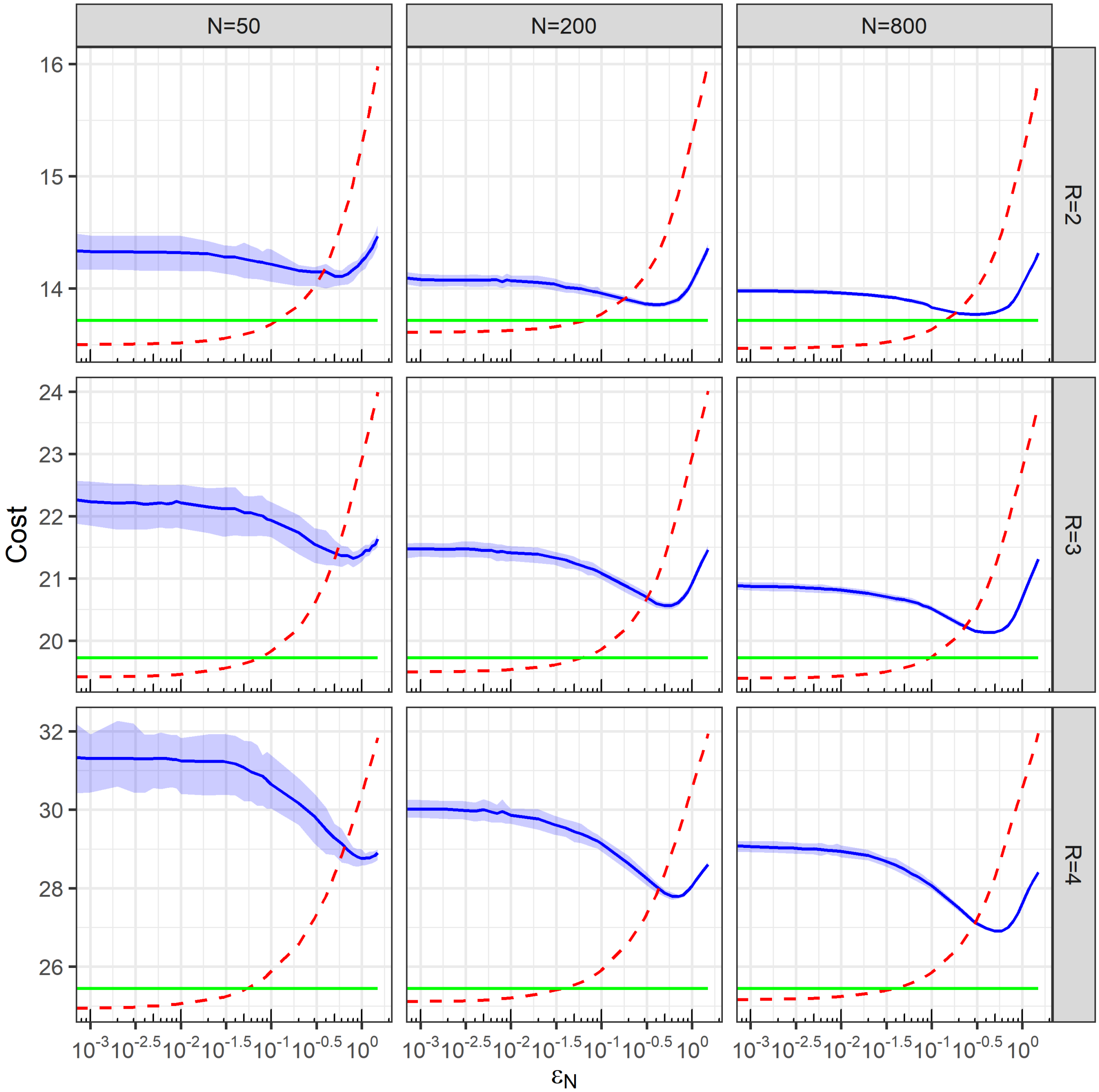
$$P^{\mathcal{A},\ell} = \frac{1}{10000} \sum_{i=1}^{10000} \mathbb{1}_{\{\mathbf{x}^{\mathcal{A},i,\ell} \in [0, \bar{\mathbf{x}}]\}},$$

Figure EC.2 Three-stage inventory replenishment problem: Histogram of replenishment decision rules for SRO-FA



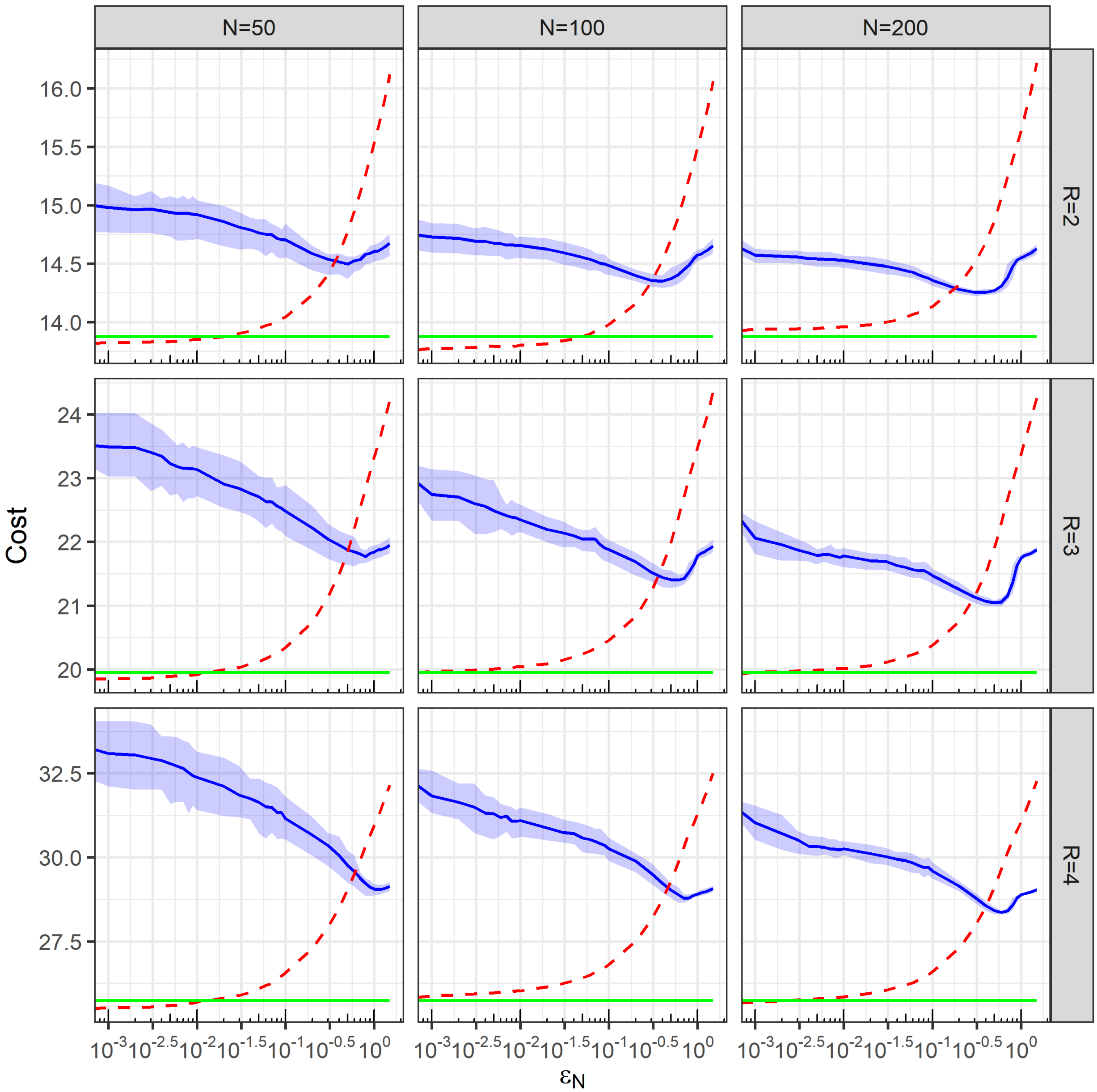
Note. The histogram corresponds to the replenishment decision rules obtained by SRO-FA in experiments where $N = 200$, $R = 3$, and $\epsilon_N = 0.7$. The light-black bars correspond to experiments in which there was no fixed cost ($f = 0$) and the dark-black bars correspond to experiments in which there was fixed cost ($f = 0.1$). For each number of retailers $0, 1, 2, 3$, the histogram shows the number of sample paths in the test set for which the replenishment policies sent a nonzero quantity of magazines to the corresponding number of retailers.

Figure EC.3 Three-stage inventory replenishment problem:
Impact of robustness parameter on SRO-FA, no fixed cost ($f = 0$)



Note. The solid black lines are the average out-of-sample cost of decision rules produced by SRO-FA, and the shaded regions are the 20th and 80th percentiles over the 50 training datasets. The dotted red lines are the average in-sample cost for SRO-FA. The green line is the benchmark cost of Problem (7). Results are shown for $N \in \{50, 200, 800\}$.

Figure EC.4 Three-stage inventory replenishment problem:
Impact of robustness parameter on SRO-FA, fixed cost ($f = 0.1$)



Note. The solid black lines are the average out-of-sample cost of decision rules produced by SRO-FA, and the shaded regions are the 20th and 80th percentiles over the 50 training datasets. The dotted red lines are the average in-sample cost for SRO-FA. The green line is the benchmark cost of Problem (7). Results are shown for $N \in \{50, 100, 200\}$.

and the *infeasibility magnitude* is approximated by

$$C^{\mathcal{A},\ell} = \frac{1}{10000} \sum_{i=1}^{10000} \min_{\mathbf{y} \in [0, \bar{\mathbf{x}}]} \left\| \mathbf{x}^{\mathcal{A},i,\ell} - \mathbf{y} \right\|_1 .$$

In other words, $P^{\mathcal{A},\ell}$ tells us how frequently the projection procedure needs to be applied, and $C^{\mathcal{A},\ell}$ captures the average number of production units which are changed due to the projection procedure.

In Tables EC.1 and EC.2, for each experiment in Section 8.3 and for each approach $\mathcal{A} \in$ {SRO-LDR, SAA-LDR}, we report the average and standard deviations for $P^{\mathcal{A},\ell}$ and $C^{\mathcal{A},\ell}$ over the 100 training datasets. For almost all choices of T, α, and N, the decision rules produced by SRO-LDR are feasible for over 93% of the sample paths in testing dataset, and their *infeasibility magnitude* is below 2 units. These results imply that the projection procedure does not significantly impact the out-of-sample cost of SRO-LDR reported in Table 1. In contrast, SAA-LDR produces decision rules which have low feasibility and high *infeasibility magnitude* when N is small. This shows, for small training datasets, that the decision rules obtained by SAA-LDR can be unreliable and require significant corrections to obtain feasible production quantities.

Table EC.1 Multi-stage stochastic inventory management: out-of-sample feasibility.

			Size of training dataset (N)			
T	α	Approach	10	25	50	100
5	0	SRO-LDR	96.3(6.6)	98.9(2.1)	99.7(0.8)	100.0(0.1)
		SAA-LDR	83.0(13.4)	96.5(4.3)	99.5(1.3)	100.0(0.1)
	0.25	SRO-LDR	93.8(7.3)	95.9(3.5)	97.3(2.3)	98.1(1.1)
		SAA-LDR	79.2(12.9)	92.3(5.3)	96.5(2.8)	98.1(1.1)
	0.5	SRO-LDR	89.7(8.6)	91.0(4.9)	91.1(3.7)	94.1(2.4)
		SAA-LDR	73.4(11.3)	85.4(4.9)	89.9(3.5)	94.0(2.3)
10	0	SRO-LDR	99.6(1.0)	100.0(0.1)	100.0(0.0)	100.0(0.0)
		SAA-LDR	61.5(24.6)	99.0(1.6)	100.0(0.1)	100.0(0.0)
	0.25	SRO-LDR	99.4(1.8)	99.9(0.3)	100.0(0.1)	100.0(0.0)
		SAA-LDR	60.2(23.9)	97.8(2.2)	99.8(0.4)	100.0(0.0)
	0.5	SRO-LDR	96.7(2.9)	97.7(1.4)	98.6(0.7)	98.9(0.3)
		SAA-LDR	57.6(22.4)	93.9(3.0)	97.7(1.2)	98.9(0.3)

Mean (standard deviation) of the percentage of the 10,000 sample paths in the testing dataset for which the linear decision rule resulted in feasible production quantities ($P^{\mathcal{A},i}$). In other words, 100% minus the above values indicates the percentage of sample paths in the testing dataset for which the production quantities needed correction. The mean and standard deviation are computed over 100 training datasets for each value of N, T, α.

Table EC.2 Multi-stage stochastic inventory management: infeasibility magnitude.

			Size of training dataset (N)			
T	α	Approach	10	25	50	100
5	0	SRO-LDR	0.5(1.6)	0.1(0.3)	0.0(0.0)	0.0(0.0)
		SAA-LDR	4.6(6.5)	0.4(0.9)	0.0(0.1)	0.0(0.0)
	0.25	SRO-LDR	0.8(1.4)	0.4(0.6)	0.2(0.3)	0.1(0.1)
		SAA-LDR	5.7(6.6)	0.9(1.1)	0.2(0.3)	0.1(0.1)
	0.5	SRO-LDR	1.7(2.1)	1.1(0.8)	1.0(0.6)	0.6(0.4)
		SAA-LDR	7.8(7.3)	2.0(1.2)	1.1(0.7)	0.6(0.4)
10	0	SRO-LDR	0.0(0.1)	0.0(0.0)	0.0(0.0)	0.0(0.0)
		SAA-LDR	218.2(1417.0)	0.1(0.2)	0.0(0.0)	0.0(0.0)
	0.25	SRO-LDR	0.0(0.2)	0.0(0.0)	0.0(0.0)	0.0(0.0)
		SAA-LDR	218.8(1417.2)	0.2(0.3)	0.0(0.0)	0.0(0.0)
	0.5	SRO-LDR	0.4(0.4)	0.2(0.2)	0.1(0.1)	0.1(0.0)
		SAA-LDR	220.3(1417.4)	0.7(0.5)	0.2(0.2)	0.1(0.0)

Mean (standard deviation) of the average *infeasibility magnitude* on the testing dataset resulting from applying the projecting procedure on the production quantity produced by the linear decision rule ($C^{\mathcal{A},i}$). The mean and standard deviation are computed over 100 training datasets for each value of N, T, α.