Smart Data-Driven Decision-Making with Uncertainty: Methods and Applications in Supply Chain Management

by

Meng Qi

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Zuo-Jun (Max) Shen, Chair
Professor Paul Grigas
Professor Phillip Kaminsky
Professor Terry Taylor

Summer 2022

Smart Data-Driven Decision-Making with Uncertainty: Methods and Applications in Supply Chain Management

Abstract

Smart Data-Driven Decision-Making with Uncertainty: Methods and Applications in Supply Chain Management

by

Meng Qi

Doctor of Philosophy in Engineering - Industrial Engineering and Operations Research

University of California, Berkeley

Professor Zuo-Jun (Max) Shen, Chair

The past decade has seen tremendous growth in the availability of voluminous high-quality data in many modern decision systems such as supply chain management. Big data provides new opportunities to tackle one of the main difficulties in decision-making systems – uncertain behavior driven by some unknown ground truth probability distribution. My dissertation mainly focuses on answering the following question:

How do we use (non-perfect) data samples generated from the unknown distribution to make more automatic and reliable decisions, especially for supply chain management and retail operations?

In Chapter 2, this dissertation study a practical multi-period inventory management problem faced by online retailers and propose a practical, data-driven, end-to-end (E2E) framework for multi-period inventory management that leverages the power of deep learning. Instead of adopting the two-step approach that first forecasts demand and vendor lead-time (VLT) separately and then, based on the predictions, solves for a solution that minimizes inventory costs, we directly learn the underlying mapping from features to the optimal solution by adopting a supervised learning approach.

In Chapter 3, we propose an integrated conditional estimation-optimization (ICEO) framework that uses statistical learning theory to estimate the underlying conditional distribution while considering the structure of the optimization problem. This method provides a generic learning framework and applies to a broad class of convex contextual stochastic optimization problems with uncertainty in the objective. We provide statistical performance guarantees and address the computational challenges that arise in performing this framework.

In Chapter 4, we focus on data-driven distributionally robust optimization (DRO) methods. We examine the conditional quantile prediction problem, which is equivalent to the

contextual newsvendor problem, and propose a DRO framework. In contrast to the traditional setting where the design points are considered as random and i.i.d., we consider a fixed-design setting, which is more suitable for cases when features are pre-designed, such as through a series of controlled experiments.

Chapter 5 continues our study on the contextual newsvendor problem and considers inter-temporal correlations and moderate non-stationarities in the contextual demand process. Under these comparatively more realistic assumptions, we provide performance guarantees in the form of out-of-sample generalization bounds.

To My Parents.

Contents

List of Figures

Acknowledgments

I would like to express my deepest appreciation to my advisor, Professor Zuo-Jun (Max) Shen, for his continuing support of my research and career development during the whole length of my PhD study. He guided me into the general field of Operations Research and taught me to study interesting and important research problems. Without his advice and encouragement, I would not have chosen a career in academia.

I am also indebted to Professor Paul Grigas for his excellent advice and incredible mentorship. We collaborated on the results in Chapter 3, and I learned many good research habits from him. I also learned teaching skills when I was his teaching assistant. I would also like to thank my other two dissertation committee members: Professor Phil Kaminsky and Professor Terry Taylor, for their great services. It is their enthusiasm and valuable feedback that made my qualifying exam and dissertation workshops thoughtful and rewarding. Furthermore, I want to thank Professor Zeyu Zheng for his support during our collaboration.

I am very grateful to have great lab mates and fellow PhD students in the program. I treasured the experience of collaborating with Ying Cao and Mengxin Wang. I would also like to thank my roommates and fellow PhD students, Renyuan Xu and Xu Rao, for their friendship. I also would like to thank all my peers in the department for the fun conversations, Lunch& Learns, and the good time we spent together. My thanks go to: Erik Berteli, Haoyang Cao, Junyu Cao, Chen Chen, Shiman Ding, Yuhao Ding, Ilgin Dogan, Pelagie Elimbi Moudio, Salar Fattahi, Han Feng, Dean Grosbard, Anran Hu, Yiduo Huang, Arman Jabbari, Titouan Jehl, Hansheng Jiang, Yusuke Kikuchi, Jiung Lee, Kevin Li, Tianyi Lin, Heyuan Liu, Alfonso Lobos, Cheng Lu, Julie Mulvaney-Kemp, Matt Olfat, Sangwoo Park, Wei Qi, Jiaying Shi, Auyon Siddiq, Quico Spaen, Birce Tezel, Mark Velednitsky, Nan Yang, Min Zhao, Mo Zhou, Ruijie Zhou. I would also like to thank my labmates in the CEE department: Yanqiao Wang, Chao Mao, Yunduan Lin, Yiduo Huang.

I would also like to thank all of my coauthors outside Cal: Professor Ho-Yin Mak, Yuanyuan Shi, Yongzhi Qi, Chenxin Ma, Rong Yuan, Di Wu.

I owe much gratitude to my family. I have to thank my parents for their unconditional love and support throughout my life. I would also like to express my appreciation and love to my husband, Tairu Lyu, who believed in me even when I did not. It is their love and support that make me the best of me. I also want to thank my cat, Lucky, who joined my life during the COVID-19 pandemic. He kept me company and entertained during the last and the most difficult year of my PhD study.

Chapter 1

Introduction

In the past decade, the greatest innovation in retailing business has arguably been digitalization. Digital technology and big data have arguably drawn the attention of practitioners and renovated their operations strategies. Figure 1.1 shows the relative Google search frequencies of three key phrases related to supply chain management over the past 10 years (compiled from Google Trends). For decades, researchers and practitioners have been employing operations research (OR) techniques in optimizing their operations. In the supply chain context, a large part of this methodology is synonymous with the phrase supply chain optimization. With digitalization, data analytics has quickly become a core element of supply chain management. Figure 1 shows that the popularity of supply chain analytics has rapidly grown, and surpassed supply chain optimization (the search frequency for which has stayed relatively flat) in around 2014 to 2015. While it can be rightfully argued that supply chain analytics is a broader concept that subsumes supply chain optimization as a subset, this observation still suggests that many (especially practitioners) now embrace data-driven (analytics) approaches for supply chain management and began to investigate operations problems from a data-driven, analytical lens.

Therefore, key technologies such as artificial intelligence (AI) plays an important role. As shown in Figure 1.1, the interest in supply chain AI has grown from practically zero a decade ago to about half the level of supply chain optimization today. Most AI, or in other words, machine learning (ML), techniques are designed for predictive analytics. Predictive analytics mainly focuses on finding correlations to make forward-looking projections from historical observations. However, predictive analytics is not enough for operations management because what we really need is decision-making. This dissertation combines tools and concepts from optimization, machine learning, and statistics to investigate methodological techniques and practical methods. The main focus of this dissertation is to answer the following question:

How do we use (non-perfect) data samples generated from the unknown distribution to

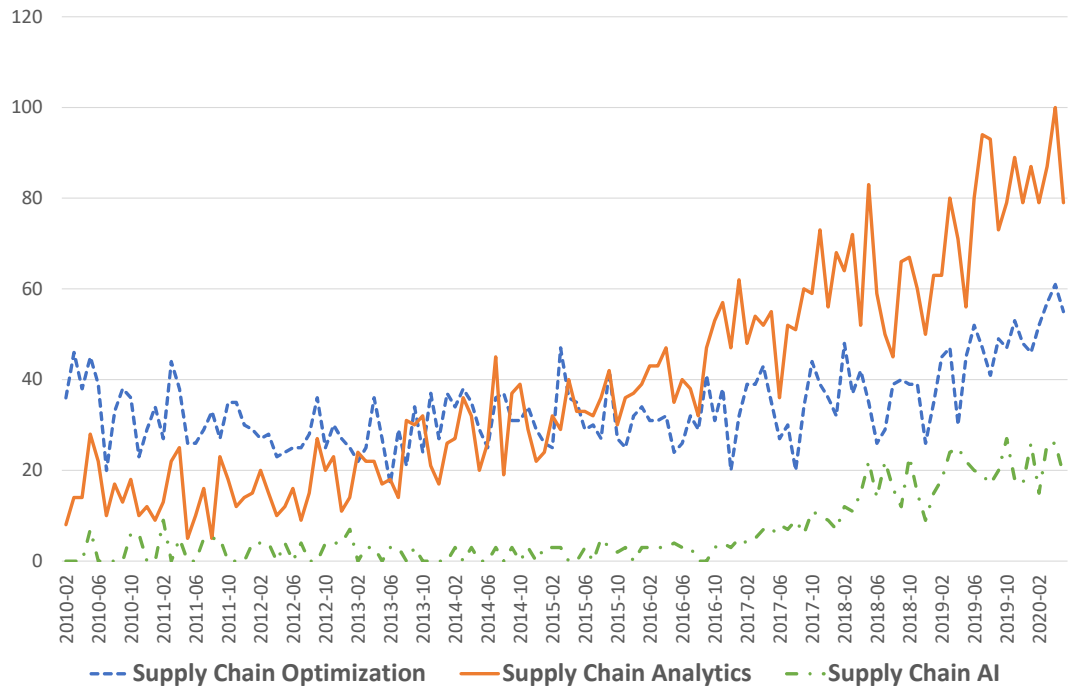


Figure 1.1: Relative Search Frequencies for Key Supply Chain Phrases on Google

make more automatic and reliable decisions, especially for supply chain management and retail operations?

To answer this question, we investigate three interconnected topics: (i) end-to-end methods that combine prediction and optimization, (ii) distributionally robust data-driven methods, and (iii) learning for decision-making when data is not independent and identically distributed (i.i.d.).

In Chapter 2, this dissertation studies a practical multi-period inventory management problem faced by online-retailers and propose a practical, data-driven, end-to-end (E2E) framework for multi-period inventory management that leverages the power of deep learning. While big data allows online retailers to offer a larger variety of products and fast delivery, it also provides new challenges in how to derive more automatic, data-driven solutions for inventory management. In this chapter, we propose a E2E model that leverages the power of deep learning. Instead of adopting the two-step approach that first forecasts demand and

vendor lead-time (VLT) separately and then, based on the predictions, solves for a solution that minimizes inventory costs, we directly learn the underlying mapping from features to the optimal solution by adopting a supervised learning approach. The main challenge is the lack of training data because the optimal solutions for historical orders are not available. We propose a labeling method that efficiently constructs the target decisions for a large-scale dataset. In numerical experiments, the E2E method outperforms multiple baseline models, including predict- then-optimize methods and common practices in industry. We implemented this algorithm at a leading online retailer for the replenishment of thousands of products and conduct a field experiment. With this method, inventory costs have been reduced by up to 25%. The content of this chapter correspond to the material in our paper [117].

In Chapter 3, I continue my research on methodologies for end-to-end decision-making. Instead of focusing on practical solutions, we model the conditional distribution of the random parameter given the contextual feature and then estimate it with an objective that aligns with the ultimate optimization problem. To the best of our knowledge, this is the first model that includes the conditional distribution, which allows us to fundamentally capture the correlation between the contextual features and the random parameter, especially when there are model misspecification and feature distribution shift. We show that our framework is asymptotically consistent under moderate regularization conditions and further provide finite-sample performance guarantees in the form of generalization bounds. With the help of an approximated solution oracle, we address the issue of non-differentiability, which is the main difficulty encountered while training end-to-end learning models. Distinguished from existing methods, our work is the first attempt to address the non-differentiability issue in optimal solution mapping with a performance guarantee. We also provide exact solution methods when the nominal optimization problem adopts a semi-algebraic objective function. The content of this chapter correspond to the material in the paper [58].

Chapter 4 focuses on data-driven distributionally robust optimization (DRO) methods. We examine the conditional quantile prediction problem, which is equivalent to the contextual newsvendor problem, and propose a DRO framework. In the existing literature, a common assumption is that the design points (features) are randomly generated in an i.i.d. manner, which is often unrealistic especially when features are pre-designed, such as through a series of controlled experiments. In contrast to the traditional setting, we consider a fixed-design setting, which is more suitable for cases when features include prices, promotions, or drag doses. As the data samples are not i.i.d. in the fixed design setting, existing DRO methods no longer apply. To overcome this difficulty, we propose a two-step regress-then-robustify method and construct the ambiguity set based on a surrogate empirical distribution of the regression residuals. We then prove the measure concentration results, which leads to finite-sample performance guarantees and asymptotic consistency. To the best of our knowledge, this is the first attempt to investigate DRO solutions with the fixed design setting. The content of this chapter correspond to the material in [115]. Together with Ying Cao, we

came up with the idea and results in this chapter (an earlier version of the results covered in Sections 4.1-4.7 are also stated in Ying's dissertation [30]).

Chapter 5 continues the study on the contextual newsvendor problem. This chapter investigates the performance guarantees for solving data-driven contextual newsvendor problems when the contextual data contain intertemporal dependence and non-stationarities. Most existing work assume that the contextual data are independent and identically distributed (i.i.d), but such assumptions are often violated in real operational systems. By accommodating these naturally-arising setting, our work adopts comparatively more realistic assumptions and provides performance guarantees in the form of out-of-sample generalization bounds. The contents of this chapter correspond to the material in our paper [116].

Chapter 2

A Practical End-to-End Inventory Management Model with Deep Learning

2.1 Background and Motivation

Inventory management has been an active research topic in management science for over a century. For comprehensive coverage on this topic, please refer to textbooks such as [156, 134]. However, in today's digital and fiercely competitive world, e-commerce companies are facing new challenges in inventory management owing to the increasing level of customer diversity, the increasing variety of products, and the higher level of service required. For example, on large e-commerce platforms (such as Amazon and JD.com), hundreds of millions of products are simultaneously sold, with various demand patterns that require different replenishment strategies. Hence, it is critical to develop a framework that would be able to identify the optimal/close-to-optimal strategy automatically for different demand, since they can't manage these many products efficiently by current practices.

Motivated by the previously mentioned requisite, we provide a framework that automatically outputs the replenishment decisions for a large number of SKUs. More specifically, we consider the multi-period inventory management problem over a finite horizon, while both demand and vendor leadtime (VLT) are considered to be stochastic. The study of this kind of problem starts from [73] and [41], where the merit of (s, S) policies and myopic base stock policies are demonstrated. However, implementing such policies requires an estimation/prediction of certain parameters/random variables then incorporating the estimation/prediction into those policies (see [141, 145, 154] as represnetatives). This type of solution paradigm where first predicting random variables of interest then incorporating the forecast results into optimization stage is the so-called predict-then-optimize (PTO) solution framework ([44]). Although being widely adopted, it decouples the prediction stage and

optimization stage. Consequently, the optimization step does not use the input data in an optimized manner and useful information can be substantially lost in the PTO process.

Instead of implementing a conventional two-step framework, we propose a data-driven, end-to-end (E2E) framework for this problem. The term "end-to-end" means training a model to output the inventory replenishment decision directly from input data without any intermediate steps. The integration of prediction and optimization has been investigated in several existing works about inventory management [10, 109]. Both [10] and [109] focus on the feature-based newsvendor problem, which is essentially a quantile regression problem of demand and a direct recipe is provided by statistical learning theory [80]. However, the multi-period replenishment problem is substantially different from newsvendor problem in two aspects: first, it adopts a multi-period setting where current decisions affect the future instead of the single-period setting in newsvendor problem; second, there are two types of uncertainties (stochastic demand and stochastic VLT), while newsvendor problem only considers one source of uncertainty (stochastic demand). Therefore, the multi-period inventory problem is an essentially more complicated problem and there is no simple closed-form solution for the optimal replenishment decision.

The lack of proper labels makes a fundamental challenge for developing an end-to-end supervised learning framework in the multi-period setting. To overcome it, we propose a labeling method by solving a dynamic programming problem and label each sample of order with the optimal decision under its realization (Theorem 1). As for the choice of model structure, we design a modular deep learning framework where we have an individual prediction block (a recurrent neural network) for demand and an individual prediction block (a multiple layer perceptron) for VLT respectively. Then these two blocks join together with other features such as review period and initial inventory, to produce the final replenishment decision output. Compared with a fully connected neural network over all features, our design reduces the computational complexity in magnitudes, while providing convenience on explanation and good performance as well. Such a modular-designed framework, together with the labeling process, could form a general recipe for developing End-to-End learning models for other large-scale supply chain management problems. We conduct a series of thorough numerical experiments that consist of both off-line comparisons and a field experiment. In off-line numerical experiments, we compare the performance of the proposed E2E model with existing PTO benchmarks using real-world datasets from JD.com. The results demonstrate that the E2E model could reduce the total inventory management cost compared with multiple PTO methods.

The E2E algorithm has been implemented at JD.com for some SKUs in "tea set" and "pastry essentials & seasoning" categories starting from February 2020. As JD.com gradually expanding the number of SKUs with E2E algorithm implemented, we conduct a systematic field experiment for 30 days from March 30, 2020 to April 30, 2020. The experiment involved 61430 orders placed in 12 distribution centers for 9308 SKUs. We compare the performance

of the E2E algorithm with that of the retailer's current replenishment algorithm. The results show that the E2E algorithm dominates the current algorithm across all performance measures. More specifically, the average holding cost, stockout cost, and total inventory cost for the treatment group are all reduced by more than 25% compared with those for the control group, while two other metrics, the turnover rate and stockout ratio, are reduced by 8.8% and 34.6%, respectively. Hypothesis tests conducted confirm that all observed reductions are statistically significant. To further verify the effect of the E2E algorithm, we adopt a difference in differences approach. The field experiment results demonstrate the immediate applicability of the E2E method at JD.com.

Our work distinguishes itself from the existing literature in the following three aspects:

1. We are the first to propose an end-to-end framework for the multi-period replenishment problem. Based on a labeling method that we proposed, our framework outputs replenishment decisions directly from input features. We believe this could be a general recipe for end-to-end models while the optimization stage is complicated.
2. The proposed end-to-end learning framework automatically decides the replenishment strategy for various demand patterns. Our approach outperforms multiple baseline models, including the current practice of a leading online retailer in both offline simulations with real-life data and the field experiment. The field experiment verifies the applicability of our algorithm for real-world inventory management in the e-commerce industry.
3. We also innovate the design of deep neural networks structure by designing two modules for demand and VLT uncertainty separately. Our design reduces the computational complexity and the number of weights in magnitudes while achieving good performance. We also perform a comprehensive sensitivity analysis in Appendix B.

The remainder of this chapter is organized as follows: In Section 2.2, we review the related literature. In Section 2.3, we state the detail of problem setting and the E2E model. In Section 2.4, we test our E2E model by conducting off-line numerical experiments with real-world data. In Section 2.5, we demonstrate the design and results of the field experiment. Finally, we conclude in Section 2.6 and propose some potential future research directions. Moreover, Appendix A gives the detailed proof of Theorem 1. Sensitivity analysis of neural network architectures, datasets, and model covariates are provided in Appendix B.

2.2 Literature Review

In this section, we first briefly review existing methods for multi-period replenishment problem and classic two-step PTO algorithms. After that, we discuss some existing literature on data-driven approaches, as well as the emerging idea of integrating prediction and

optimization for inventory management problems, which is mainly based on the Newsvendor setting. Finally, we introduce the development and applications of end-to-end approaches in other fields, which serves as the incubator for the proposed E2E inventory management framework.

Multi-period inventory management problem has been studied in decades since [73]. Using dynamic programming, [73, 41] established the optimality conditions for base stock and (s,S) policies under finite and infinite horizons, respectively. Although the optimality of base stock policies has been proven under different settings ([65, 54, 105]), calculating the optimal parameter of such policy remains computationally intractable under general cases [92]. In order to get computationally tractable algorithms, one option is to use approximation algorithms. For example, [92] provided 2-approximation algorithms using dual-balancing techniques. The other option is to use heuristic algorithms such as myopic policies (see [142, 65, 64] as representative works).

However, all these aforementioned policies assume certain knowledge of demand and VLT (e.g., demand and VLT follow certain distributional models and the parameters of the model are known). In practice, such information is often unveiled to decision-makers. Therefore, a two-step predict-then-optimize (PTO) framework is widely adopted in industry for inventory management. The PTO framework first forecasts demand and VLT then incorporates the prediction into certain decision rules such as base stock and (s, S) policies stated above.

In the stage of prediction, there are two different types of forecasting methods for demand and VLT. The output of forecasting can be a point estimator or a distribution of the random variable. The first type of forecasting is widely adopted in industry since in some cases accurate prediction can be achieved by machine learning models (e.g., [53]). However, point estimation of random variables can lead to information loss, which affects the subsequent optimization stage. In contrast, if we can make perfect forecast of the random variable distribution, we have all the information in order to solve the following stochastic optimization problem. Recently, there are works that develop distributional/probabilistic forecasting models. When the random variable does not depend on external features, the distribution can be fitted by simply using empirical distribution of historical observations or by kernel density estimation ([133]). When the random variable depends on covariates, distributional/probabilistic forecasting becomes more difficult. A recent work by [18] uses a non-parametric method to approximate the distribution of the random variable conditioned on covariates by weighted empirical distribution. However, this benchmark is not applicable to numerical experiments using real-world dataset that includes time series features, due to computationally difficulty. [24] forecast multiple quantiles of demand to gain more distributional information. Their approach is similar to the benchmark BM1 in our offline numerical experiments. [6] propose another nonparametric method for conditional density estimation using kernel mixture network. In their work, densities are assumed as linear combinations of a family of kernel functions and the weights are determined by a deep neural network.

Therefore, it requires a lot of computational effort thus not applicable in our setting. Moreover, we want to highlight the fact that, due to the multi-period setting in our problem, even though we have a practical method for estimating the conditional distribution, solving corresponding stochastic dynamic programming is also computationally difficult (we refer to [92] for a more detailed review).

Since the traditional two-step approaches that separate the prediction from the optimization often lead to sub-optimality, there has been a trend to perform these two steps simultaneously in the recent literature on data-driven inventory management [10, 109, 96, 36, 18]. Such integration attempts could be realized through operational statistics. [96] demonstrated the existence of improved operational statistics (in contrast to the use of the maximum likelihood estimator) by integrating the prediction and optimization steps on several demand distributions; [36] further studied how to obtain the optimal operational statistics in a Bayesian framework. Both [96] and [36] only studied the Newsvendor problem.

[10, 109] investigated the concept of integration in the feature-based Newsvendor situation, where one has access to past demand observations, as well as to a large number of related features. [10] studied this problem with the Newsvendor loss function and assumed a linear relationship between the features and the Newsvendor quantile (i.e. the solution). They analytically showed that their approach can perform better than the SAA and the separated estimation and optimization method. [109] adopted a multiple-layer perceptron model that optimized the order quantity. However, all the aforementioned works have been focused on the Newsvendor problem, in which neither the connection between the periods nor the vendor lead time has been considered, which are both important factors in practice.

Beyond inventory problems, [18, 44] studied the "integration" philosophy for general optimization problems. [18] combined machine learning and optimization techniques for decision-making purposes when features (referred to as auxiliary quantities in the paper) were available. Their idea was to construct weight functions from data through machine-learning methods and to incorporate these weights to the objective in the optimization procedure. Under the context of linear programming, [44] proposed a "smart PTO" framework that directly leveraged the optimization problem structure forming the loss function.

Another branch of research that served as an impetus for our work originated from the machine-learning community. In recent years, there has been a dramatic increase in the number of systems built on "E2E learning" [39]. This term refers to a learning framework, the ultimate goal of which is directly predicted from raw inputs rather than from intermediate steps. This concept has been successfully applied to a wide range of tasks, such as finance [14], image recognition [144] and robotics manipulation [93]. These lines of works have provided certain valuable inputs to us in terms of integrating prediction and optimization; however, such an "E2E" approach has not yet been studied with a focus on the general supply-chain management problem.

2.3 Model

In this section, we first describe the multi-period replenishment problem; In Section 2.3.1, this problem is presented with a dynamic programming framework. In Section 2.3.2, we explain our E2E model with emphasis on its two key components, the property of the optimal dynamic programming solution and the deep learning network structure.

2.3.1 The Multi-Period Replenishment Problem

In this work, we consider the multi-period inventory management problem with stochastic demand and VLT. The details are as follows: for a single item at a single location, we consider a finite horizon of discrete periods $1, \ldots, T$, where T is the end of the horizon. Over the T periods, there is a sequence of random demands, denoted by $D_t, \forall t = 1, \ldots T$. Let I_t denote the inventory level at the beginning of period t. The inventory level can be positive if we have inventory excess on hand or negative if we have an inventory shortage and, hence, backorders. As a result, at the end of each period, we either incur a holding cost of h for each excess unit or a stock-out cost of b for each back-ordered unit.

We consider periodic review policies and assume that review periods are given as a sequence of dates. That is, we assume there are totally M orders from period 1 to T that are placed at $t_m, \forall m = 1, \ldots, M$. This assumption is aligned with the real-world practice, where a fixed schedule is typically held for order placement (e.g., one can place orders on Tuesdays and Fridays). In this problem, we consider the stochastic VLT. That is, the m^{th} order placed at period t arrives at period $t + L$, where L is a random variable that only takes positive integer values. Hence, the arrival time of orders, denoted by $v_m = t_m + L_m$, $\forall m = 1, \ldots, M$, are also random variables. Moreover, we assume there are no crossing-over of order arrivals. Although some of the aforementioned assumptions may be unrepresentative (such as back-order and no crossing-overs), in section 2.4.2 we relax some of them and test our proposed method under a more realistic setting.

Hereafter, D_t and L_m will denote the random variables; d_t and l_m denote the realization of demand at period t and the realization of VLT of the m^{th} order, respectively. At each period, the system first updates the inventory level by checking if any order has arrived; then, demand occurs. Let a_m denote the order quantity for the m^{th} order. At the end of the period, either a holding cost or back-order cost occurs. The inventory level updates follow the equation below,

$$I_{t+1} = I_t - D_t + \sum_{m=1}^{M} a_m \mathbb{1}\{t = t_m + L_m\}. \tag{2.1}$$

Let the cost that occurred at period t be denoted by S_t, we have

$$S_t = h[I_t - D_t + \sum_{m=1}^{M} a_m \mathbb{1}\{t = t_m + L_m\}]^+ + b[-I_t + D_t - \sum_{m=1}^{M} a_m \mathbb{1}\{t = t_m + L_m\}]^+. \quad (2.2)$$

where $[\cdot]^+$ denotes $\max\{\cdot, 0\}$.

Our aim is to minimize the expected cost during the finite horizon by choosing the order quantities at given periods, that is

$$\min_{a_1,\ldots,a_M} \mathbb{E}\left[\sum_{t=1}^{T} S_t\right]. \quad (2.3)$$

S_t is defined by (2.2) and the updates of I_t follow (2.1). Note that the expectation is taken over the joint distribution of the demand $\{D_t\}_{t=1}^T$ and the VLT $\{L_m\}_{m=1}^M$.

2.3.2 End-to-End (E2E) Model

The goal of the replenishment problem is to determine the best order quantity, $a : f(\mathbf{x}) \in R$ at each given review point, after having observed all the features, $\mathbf{x}$. Such features can include historical demand, VLT, item specifications, and temporal information (day, month, season).

To find the mapping function $f(\cdot)$, we first train the model with historical data. For each historical replenishment time point, with observed feature vector $\mathbf{x_i}$, we need to compute a_i^*, which is the corresponding optimal order quantity. This step is referred to as "labeling" for supervised learning algorithms (Section 2.1.4 [66]). We will describe the details of the labeling method in Section 2.3.2.1. When the associated label for each observation is completed, we can establish the mapping with the following training objective:

$$\min_{f:\mathcal{X}\to R} \sum_{i=1}^{N} L(f(\mathbf{x}_i); a_i^*), \quad (2.4)$$

where N is the total number of training data, L is the loss function that is defined based on the difference between the model prediction $f(\mathbf{x}_i)$ and the optimal order quantity a_i^*. In particular, we consider neural network models for function f and we will describe the details of the neural-network structure in Section 2.3.2.2.

2.3.2.1 Labeling the optimal order quantity

Unlike the newsvendor problem, for which the optimal solution is the $\frac{b}{b+h}$ quantile of demand distribution, the optimal solution of the multi-period inventory problem is not

straightforward to calculate. In this section, we will analyze the properties of the optimal order solution, hence producing labels $a_i^*, i = 1, ..., N$ for the training set.

Given the order place and arrival time, and the demand at every time step, we can compute the optimal quantity for each order using the dynamic programming framework. It should be noted that within periods $1, \ldots, T$, there are M orders placed. Moreover, $t_m \in \{1, \ldots, T\}, m = 1, \ldots, M$ denotes the time period in which the m^{th} order is placed (the quantity can be 0). In a similar manner, let v_m denote the time when the m^{th} order arrives. It should be stressed that we assume no crossover of lead time. With given demand, we can formulate the recursion as

$$V_m(I_{v_m}) = \min_{a_m \geq 0} \sum_{s=v_m}^{v_{m+1}-1} h[I_{v_m} + a_m - d_{[v_m,s]}]^+ + b[d_{[v_m,s]} - I_{v_m} - a_m]^+ + V_{m+1}(I_{v_m} + a_m - d_{[v_m,v_{m+1}-1]}), \tag{2.5}$$

where $V_m(I_{v_m})$ is the optimal cost over interval $[v_m, v_{m+1} - 1]$, $d_{[i,j]} := \sum_{t=i}^{j} d_t$.

The following theorem describes the closed-form solution of (2.5); hence, it provides an efficient approach to label the training and testing data set.

Theorem 2.3.1. *The optimal multi-period inventory replenishment problem described by (2.5) is decomposable, i.e.,* $a_m^{**} := \arg\min_{a_m \geq 0} \left\{ \sum_{s=v_m}^{v_{m+1}-1} h[I_{v_m} + a_m - d_{[v_m,s]}]^+ + b[d_{[v_m,s]} - I_{v_m} - a_m]^+ + V_{m+1}(I_{v_m} + a_m - d_{[v_m,v_{m+1}-1]}) \right\} = \arg\min_{a_m \geq 0} \sum_{s=v_m}^{v_{m+1}-1} h[I_{v_m} + a_m - d_{[v_m,s]}]^+ + b[d_{[v_m,s]} - I_{v_m} - a_m]^+$. *In addition, the closed form of the optimal solution is* $a_m^{**} = \max\{d_{[v_m,s^*]} - I_{v_m}, 0\}$, *where* $s^* = \lfloor \frac{b(v_{m+1}-v_m)}{h+b} \rfloor + v_m$.

Remark. *Theorem 2.3.1 indicates that labeling using ex-post optimal order quantities is practical, i.e. computationally efficient for large dataset. If, instead of deterministic dynamic programs, stochastic dynamic programs are solved to get labels, the labeling process becomes computationally intractable.*

2.3.2.2 Neural-network structure

When the associated labels for training data are obtained, we train a neural-network model f by solving the optimization problem in (2.4). The general structure of the neural-network model is shown in Figure 2.1.

The inputs of the E2E model include five parts, where `Input_DF` and `Input_VLT` represent features related to demand and VLT, respectively; `Input_basic` is the set of general item-level features, such as product categories, warehouse locations, and brand names. The remaining two features, review period and initial stock level, are directly fed into one of the hidden layers because they are not intended to generate any cross terms with other features. The E2E model has three outputs, where the main output `Out1` $\in \mathbb{R}$ represents the final

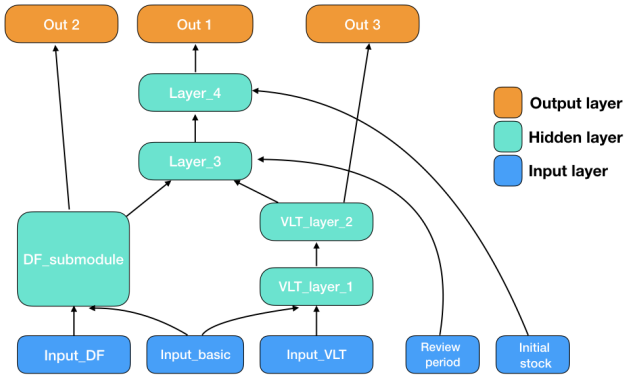


Figure 2.1: Neural network structure of the E2E Model.

replenishment decision. In addition, there are two accessory outputs `Out2` as the demand forecast and `Out3` as the VLT forecast.

All hidden layers, except for `DF_submodule`, are fully connected layers with rectified linear unit (ReLU) activation function and dropout layers [136] to prevent overfitting. The `DF_submodule` is designed as a multi-quantile RNN (MQRNN), which receives multiple time series (e.g., demand time series, promotion time series) as inputs and produces a daily demand prediction over a set of quantiles as outputs. We use MQRNN because of its demonstrated performance in demand forecasting in the e-commerce industry [148, 47].

The training objective function is defined as

$$\min_{\theta} \sum_{i=1}^{N} \left\{ L(\texttt{Out1}_i; a_i^*) + \lambda_1 \hat{L}_1(\texttt{Out2}_i, a_{DF,i}^*) + \lambda_2 \hat{L}_2(\texttt{Out3}_i, a_{VLT,i}^*) \right\}, \tag{2.6}$$

where θ is the set of neural network parameters to be optimized, N is the total number of training data, and λ_1, λ_2 are two small positive constants penalizing the demand and VLT prediction error. A few reasons lead us to include three terms in (2.6), rather than only the first term. First, the optimization of the two forecasting outputs (i.e., `Out2` as the demand forecast and `Out3` as the VLT forecast) act as a guide for faster model training. Without them, it becomes more difficult for each submodule of the network to be trained as desired. Moreover, `Out2` and `Out3` can be used for monitoring the model performance. While running the E2E model in real-time, the observation of any anomalies in these two outputs can help decision-makers detect any abnormal output on replenishment amounts and analyze the reason behind it.

The above network structure is designed with the knowledge that the replenishment decision would be made using the information of the demand and the VLT, whose feature sets hardly overlap and are barely related. Thus, compared with a fully connected network over all features, our design reduces the computational complexity and the number of weights in magnitudes, while providing convenience on explanation and good performance as well.

2.4 Numerical Experiment

In this section, we test the performance of E2E model using real-world data. In Section 2.4.1, we introduce several common two-step predict-then-optimize (PTO) benchmarks. In Section 2.4.2, we evaluate the performance of E2E model with different PTO benchmarks, and highlight the benefits of both the one-step decision framework and the deep learning architecture. In all experiments, we use real data from JD.com, one of the largest online retailers in China. It owns hundreds of warehouses to manage inventory and has replenishment agreements with tens of thousands of vendors. All neural networks are implemented in PyTorch [111] and trained on a server with an NVIDIA Tesla P40 GPU. We refer to Appendix B for more details of model parameters and a sensitivity analysis.

2.4.1 Comparison with Two-Step Replenishment Methods

In order to show the performance of our E2E model, we compare the E2E model with several currently in-use two-step PTO methods.

We start with two widely adopted base stock policies. In "Normal" base stock, the daily demand is assumed to be independent and identically distributed (i.i.d.) and follows the Normal distribution. Thus, the base-stock level is computed as

$$BM_{normal} = \mu_D(R + \mu_{VLT}) + \phi^{-1}(\frac{b}{b+h})\sqrt{(R + \mu_{VLT})\sigma_D^2 + \mu_D^2\sigma_{VLT}^2}, \tag{2.7}$$

where R is the review period. The mean (μ_D, μ_{VLT}) and the standard deviation (σ_D, σ_{VLT}) are estimated using historical data that contains the demand and the VLT of the past 180 days. Similarly, in "Gamma", daily demand is assumed to be i.i.d. and follows Gamma distribution. Hence, the sum of $(R+\mu_{VLT})$ days of demand, denoted by $\bar{D}$, follows $Gamma((R + \mu_{VLT})k, \theta)$, where θ and k are estimated using the demand data of the past 180 days. The base-stock level is computed as

$$BM_{gamma} = Q_{\bar{D}}^{gamma}(\frac{b}{b+h}), \tag{2.8}$$

where $Q_{\bar{D}}$ denotes the quantile function of $\bar{D}$.

In addition to the two base-stock policies, we also want to compare the E2E model with PTO benchmarks. As mentioned earlier in Section 2.2, there are two different types

forecasting in PTO method: one type of PTO method estimate a point prediction of the random variable while the other type predicts the distribution of the random variable. If a PTO method succeeds to achieve perfect prediction of the joint conditional distribution of demand and VLT, then theoretically, (2.3) can be solved to optimality. However, as we reviewed in Section 2.2, there barely exists applicable method for our problem setting. Moreover, even with reliable forecastings of the joint distribution, we still need to solve a corresponding stochastic dynamic programming problem, which is computationally intractable.

Therefore, we construct the following two PTO benchmarks using a MQRNN for demand forecasting and a MLP for VLT prediction for fair comparison.

1) **BM1.** First, we let $\hat{d}_m = \sum_{t=t_m}^{v_{m+1}-1} d_t$ denote the total demand within two adjacent order-arrival times, namely t_m and v_{m+1}. Note that v_m can be calculated based on VLT prediction. Then the $b/(b+h)$ quantile of $\hat{d}_m$ is predicted, equivalent to solve the following problem:

$$\min_{a_m \geq 0} \mathbb{E}_{\hat{d}_m}[b(\hat{d}_m - a_m - I_{t_m})^+ + h(a_m + I_{t_m} - \hat{d}_m)^+], \tag{2.9}$$

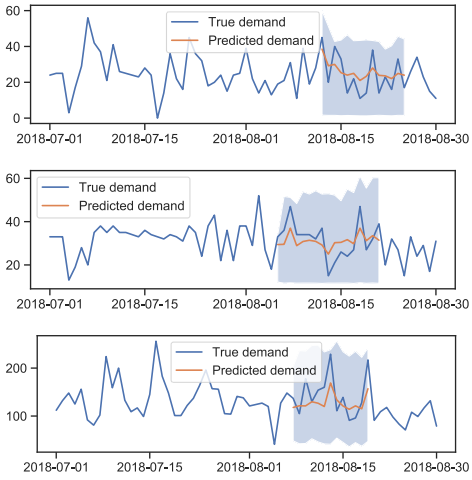
BM1 then can be considered as a PTO method with point prediction of VLT and distributional forecasting of demand. In the optimization stage, the multi-period problem is approximated by a single-period Newsvendor problem.

2) **BM2.** An alternative PTO method is to first sequentially forecast the future demand within two adjacent order-arrival times, that is $d_{t_m}, d_{t_m+1}, ..., d_{v_{m+1}-1}$, and then calculate the optimal decision, a_m, by minimizing the following accumulated inventory cost

$$\min_{a_m \geq 0} \sum_{t=v_m}^{v_{m+1}-1} h[I_{t_m} + a_m - d_{[t_m,t]}]^+ + b[d_{[t_m,t]} - I_{t_m} - a_m]^+. \tag{2.10}$$

where v_m is estimated as $\hat{v}_m = t_m + \hat{l}_m$, and v_{m+1} is estimated as $\hat{v}_{m+1} = t_{m+1} + \hat{l}_m$ as well. Notice that $\hat{l}_m$ comes from `Out3` as the VLT forecast and d_t is the demand forecast. BM2 can be viewed as a PTO method with point prediction for both VLT and demand, but in optimization stage, the multi-period problem setting is kept.

The minimizers of (2.9) and (2.10) are denoted by $a^*_{Bm1,m}$ (BM1) and $a^*_{Bm2,m}$ (BM2), respectively. Two benchmarks are employed because each of them emphasizes on different stages as a two-step model. First, because predicting the total demand within a time interval is more likely to reach the desired accuracy than predicting a sequence of demands for each day, the first benchmark could yield better results in the prediction stage. However, by noticing that the objective in (2.9) contains a newsvendor loss rather than the one in our multi-period setting, one can expect that the second benchmark would yield results of higher accuracy in the optimization stage.



Method	Total cost	Holding cost	Stockout cost
OPT	3022.60	1925.70	1096.91
E2E_RNN	**3766.52** (+24.6%)	**2689.02**	**1077.50**
BM1	4207.09 (+39.2%)	2502.55	1704.54
BM2	4157.03 (+37.5%)	2254.99	1902.04
Normal	4576.05 (+51.4%)	3369.21	1206.84
Gamma	4476.17 (+48.1%)	2821.87	1654.30
E2E_GBM	4017.82 (+32.9%)	2109.93	1907.89

Figure 2.2: Left: demand forecast for three example SKUs via MQRNN. Blue line is the ground truth, 10% and 90% quantile forecasts are the lower and upper boundary of the forecast band, and 50% (median) is the orange line within the band. The shaded area is the forecast within the current order place time and next order arrival time. Right: average inventory cost under different E2E and PTO polices over 30 days per SKU.

2.4.2 Results

We test the performance of E2E method and the four PTO policies using real-world data, a 24,333 SKUs dataset under the Food & Snack Category. The input vector for each replenishment sample contains SKU profile features, daily sales, and historical VLT, which is a 132-dimensional vector. The entire dataset are split into a training and a testing dataset according to the creation date of each sample. We use the first 30-day replenishment-order data as the training and validation set, where 80% of the data are used for training and the remaining 20% are used for validation. The validation set is used to evaluate the performance of the neural network model for different combinations of hyper-parameter values, which helps to choose the optimal hyperparameters and prevent over-fitting. The remaining 30 days of data for all SKUs serves as the testing set. The performance of E2E model and benchmarks are evaluated by total inventory management cost, holding cost and stockout cost, defined via (2.2). By default, the values of b and h were set to 9 and 1, respectively.

The total inventory management cost, holding cost and stockout cost of different models over the test period are listed in Figure 2.2 (right). The averaging holding cost and stockout cost were computed as in (2.2). "OPT" refers to the optimal replenishment decisions obtained by solving the replenishment problem with known demand and VLT, which is the same approach that is used for the labeling of the data, as described in Section 2.3.2.1. For the two PTO methods "BM1" and "BM2", we use the same neural network architecture as

the E2E model, that is an MQRNN[1] for demand prediction and a 2-layer feedforward neural network for the VLT prediction. In this experiment, the MQRNN model outputs $Q = 6$ quantile prediction, including mean and quantile levels at 10%, 60%, 70%, 80%, 90% and 95%. The results for BM1 and BM2 are reported based on the best quantile with the lowest total cost.

Figure 2.2 shows that the proposed E2E deep learning model is the best compared to all benchmarks. The advantages of end-to-end deep learning model are two-fold. On one hand, the benefit of using end-to-end framework rather two-step PTO framework can be observed by comparing the cost of E2E_RNN against BM1 and BM2. Since all three algorithms use the same deep learning structure for demand and VLT prediction, while E2E_RNN adopts one-step decision framework and BM1, BM2 use two-step framework. We suspect that the prediction errors of demand and lead time compound in the optimization stage. To avoid the accumulation of error, the end-to-end framework shortens the decision process while targeting the ultimate optimization goal. On the other hand, the benefit of deep learning model versus other statistical learning models can be observed when comparing the E2E_RNN cost and E2E_GBM cost. E2E_GBM denotes the performance of the end-to-end LightGBM [75] model, which is a decision tree based algorithm that widely used in industry. Since the tree-based models are not able to process time-series features (e.g., historical demand series), we use statistical summary of demand as features including the mean, standard deviation and temporal differences. Both E2E_RNN cost and E2E_GBM algorithms use the end-to-end decision framework, while the deep learning model has better representation capacity and RNN is more powerful in modeling time series data.

2.5 Field Experiment

The proposed E2E algorithm has been implemented in JD.com since February, 2020. As of October 2020, there are more than 7,000 SKUs adopts the E2E algorithm for inventory management, showing significant inventory cost saving in production. JD.com is also working on expanding the E2E method to more categories.

In this section, we demonstrate the design and results of a 30 days field experiment conducted in JD.com's logistics system from March 30, 2020 to April 30, 2020.

2.5.1 Overview of JD.com's Auto-Replenishment System

JD.com maintains a logistics network in China that consists of about 500 Distribution Centers (DC) national-wide. Each DC manages its inventory using JD's inventory replenishment system. JD's current inventory replenishment algorithm can be viewed as a two-step (PTO) decision-making process empowered by machine learning techniques and industry ex-

[1] Predicting the total demand for Benchmark 1 and predicting the daily demand for Benchmark 2.

pertise. In the first step, the demand and VLT are predicted using state-of-the-art machine learning methods considering seasonality, geographic effect, SKU and vendor heterogeneity. [47] describes an effort to the retailer's advanced demand forecast using deep learning techniques. In the second step, the inventory replenishment decision is made based on the predictions from the first step. Service level used in current practice of JD.com is decided by a hyper-parameter called "critical ratio" which is consistent with the hyper-parameters b and h in E2E model. Generally speaking, the critical ratio for different product category doesn't have to be the same.

In JD's replenishment system, the performance of a replenishment algorithm is quantified by five key metrics: stockout rate, turnover rate and the three metrics we used in Section 2.4.2 including the total inventory management cost, holding cost and stockout cost. The stockout rate is defined as the percentage of days that stockout occurs during the experimental period, i.e. it measures the frequency of stockouts. The inventory turnover rate is calculated by dividing the average inventory level of each day by the average demand.

2.5.2 Experiment Design

To test whether the proposed E2E algorithm leads to better replenishment decisions, we conducted a field experiment during a 30 days period, from March 30, 2020 to April 30, 2020. The experiment involved 61430 orders placed in 12 DCs for 9308 SKUs. The SKUs are from 18 third-level categories, which belong to two second-level categories, "tea set" and "pastry essentials & seasoning". The details of the categories that are involved in the field experiment are listed in Table 2.1.

Table 2.1: Details of Categories Involved in the Field Experiments

Second level categories	Tea set	Pastry essentials & Seasoning
Third level categories	Tea set combination	Baking supplies
	Tea cup	Flour
	Tea kettle	Mixed-grains
	Tea tray	Rice
	Tea can	Oil
	Tea bowl	Seasonings
	Tea accessories	Dry foods
	Tea-ware decoration	Convenience foods
	Coffee set	
	Tea travel set	

An E2E model is trained for the second-level category, tea set, and for each third-level category in the second-level category, pastry essentials & seasoning. In this experiment, we use $h = 12$ and $b = 88$, which is consistent with the critical ratio in these categories. The field

experiment contains all replenishment orders placed after March 30, 2020 for each (SKU, DC) pair in the aforementioned categories and twelve DCs. A treatment group involves 1052 SKUs and 6097 (SKU, DC) pairs in total, are selected. The replenishment decisions of these (SKU, DC) pairs are made according to the proposed E2E algorithm starting from March 30, 2020. For the remaining (SKU, DC) pairs, replenishment decisions are made following JD's current algorithm. The treatment group of (SKU, DC) pairs was chosen by JD.com's managers based on business criteria. We select a control group that also include 6097 (SKU, DC) pairs using propensity score matching in order to reduce the possibility of selection bias in this process (with details stated in Section 2.5.2.1). We collect daily inventory levels, daily sales, replenishment order placement and arrival dates for all (SKU, DC) pairs in the field experiment to calculate the performance metrics including the average holding cost, average stockout cost, average total cost, average turnover rate and average stockout rate.

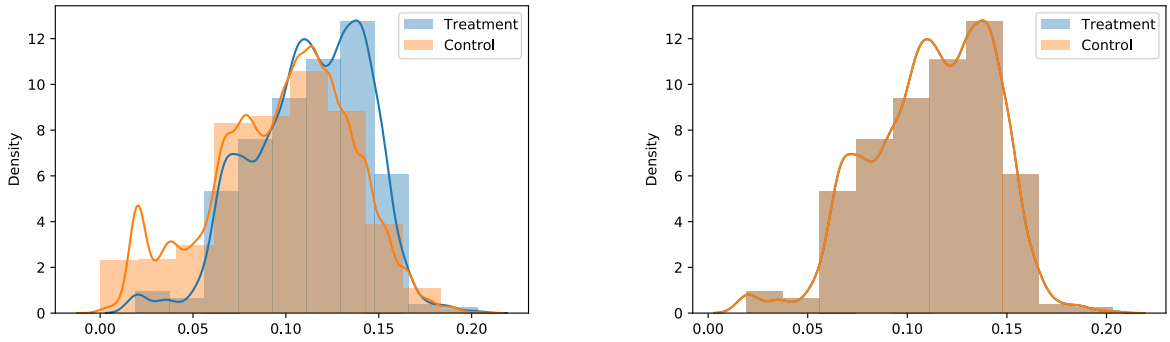
Moreover, JD's logistics system adopts less restrictive setting compared to Section 2.3.2. Instead of fully back-order, demand that can not be fulfilled by current inventory will be considered as back-ordered only if it can be fulfilled by open purchased orders. (Open purchase orders refer to those orders that have been placed but haven't arrived, in other words, replenishment that is on the way.) If it can not be fulfilled by open purchased orders, it will be considered as lost sales. In addition, there may be cross-over of orders, i.e. orders may not arrive in the same sequence as they are placed. Therefore, the field experiment can test the performance of the proposed E2E algorithm in real-world setting.

2.5.2.1 Propensity score matching

As explained earlier, the treatment group contains 6097 (SKU, DC) pairs, selected based on voluntary response, and all remaining (SKU, DC) pairs from the categories listed in Table 2.1 are the candidate control group. In order to address the issue of potential selection bias, we choose (SKU, DC) pairs in the control group by using propensity score matching (see [123] and [125] for references). For propensity score matching, we use demand and VLT as cofounder variables. Figure 2.3 visualizes the propensity score of control and treatment sets before and after matching. After matching, we have a control group with same size as the treatment group.

2.5.3 Results

In this subsection, we first compare the performance of the two algorithms during the test period using a two-sample t-test. Then to further adjust the potential differences in the treatment and control groups, we train linear regression models for each outcome metric and check the significance of the coefficients. Moreover, we evaluate the performance of both treatment and control groups before the test period and apply the difference in differences technique to further confirm the effect of applying the E2E method.



(a) Propensity score before matching

(b) Propensity score after matching

Figure 2.3: Propensity Score Matching. The overlap of the histogram and density plot in Fig (2.3b) demonstrates that the propensity score has been perfectly matched.

2.5.3.1 T-test results

Table 2.2: Comparison of Performance of Two Algorithms: t-test Results

	Holding Cost	Stockout Cost	Total Cost	Turnover Rate	Stockout Ratio
Treatment Group	598.20	496.34	1094.54	18.59	0.17
Control Group	809.54	1027.93	1837.47	20.39	0.26
Difference	**-211.34(-26.1%)**	**-531.59 (-51.7%)**	**-742.93(-40.4%)**	**1.8(-8.8%)**	**0.09(-34.6%)**
t-test p-value	< 0.001****	< 0.001****	< 0.001****	0.0102**	< 0.001****

Note: $*p < 0.1$; $**p < 0.05$; $***p < 0.01$; $****p < 0.001$.

Table 2.2 demonstrates the results of t-test for the comparison of five performance metrics: holding cost, stockout cost, total inventory cost, turnover rate, and stockout ratio. The E2E algorithm significantly reduces all five metrics with four out of five t-test p-values < 0.001 and one p-value < 0.05. In particular, the E2E algorithm can reduce the average holding cost by 26.1% and average stockout cost by 51.7%, compared to JD.com's current replenishment method. Not surprisingly, the total cost is reduced by 40.4%. The average turnover rate has also been reduced by 8.8% and the stockout ratio reduced by 34.6%.

2.5.3.2 Linear regression

Furthermore, one may have the concern that the treatment group and the control group having slightly different average demand and average VLT might lead to unreliable t-tests. To address this concern, we further consider the following linear regression model for the

outcomes of each metric,

$$\text{Outcome} = \theta_0 + \theta_1 \text{Is-E2E} + \theta_2 \text{Ave-Demand} + \theta_3 \text{Ave-VLT}, \tag{2.11}$$

where Is-E2E is a binary independent variable that represents if a SKU orders using E2E algorithm, Ave-Demand is an independent variable that represents the average demand of a SKU and similarly Ave-VLT represents the average VLT of a SKU.

The linear regression model plays a similar role as the t-tests. It aims to provide a better comparison of the treatment and control groups' average outcome by considering covariates that may affect the outcome of the metrics. Table 2.3 demonstrates the coefficients and p-values of Is-E2E.

Table 2.3: Comparison of Performance of Two Algorithms: Linear Regression

	Holding Cost	Stockout Cost	Total Cost	Turnover Rate	Stockout Ratio
Coefficient	-211.33	-531.53	-742.86	-1.16	-0.09
t-test p-value	< 0.001****	< 0.001****	< 0.001****	0.007***	< 0.001****

Note: $*p < 0.1$; $**p < 0.05$; $***p < 0.01$; $****p < 0.001$.

Table 2.3 indicates that using the E2E algorithm leads to significant reductions on all five metrics - holding cost, stockout cost, total cost, turnover rate and stockout ratio.

2.5.3.3 Difference-in-differences estimation

To further study the effect of E2E algorithm on all five metrics, we consider a difference in differences (DiD) approach. To be more specific, we evaluate the performance of the treatment and control groups from February 29, 2020 to March 29, 2020. We denote this period as the pre-experiment period. Then compare the performance with that of the post-experiment test period, March 30, 2020 to April 30, 2020. During the pre-experiment period, both treatment and control group adopt JD's current replenish algorithm and during the post-experiment period, the treatment group adopt the E2E algorithm while the control group still follows JD's algorithm.

Table 2.4 demonstrates the DiD comparison of the effect of E2E algorithm implementation. In Table 2.4, the terms "Pre-Exp" and "Post-Exp" denote the pre-experiment and post experiment periods, respectively. The results indicate that the implementation of our proposed E2E algorithm improves all five metrics. The readers may notice that, in the control group, there is a slight increment of the post-experiment metrics compared to the pre-experiment metrics. Our conjecture is that, as the economics in China begins to recover in March 2020, both demand and supply have a larger scale in April compared to March.

Table 2.4: Difference-in-Differences Estimation of E2E algorithm

	Treatment Group			Control Group				
	Pre-Exp	Post-Exp	Change	Pre-Exp	Post-Exp	Change	**DiD**	t-test p-value
Holding Cost	688.13	598.20	-89.93	786.68	809.54	22.86	**-112.79**	< 0.001****
Stockout Cost	880.50	496.34	-384.16	887.64	1027.93	140.29	**-524.45**	< 0.001****
Total Cost	1568.63	1094.54	-474.09	1674.32	1837.47	163.15	**-637.24**	< 0.001****
Turnover Rate	16.41	18.59	2.18	16.40	20.39	3.99	**-1.81**	0.041**
Stockout Rate	0.24	0.17	-0.07	0.22	0.26	0.04	**-0.11**	0.044**

Note: $*p < 0.1$; $**p < 0.05$; $***p < 0.01$; $****p < 0.001$.

2.6 Conclusions

In this work, we propose an E2E framework with deep learning models for multi-period replenishment problems, without prior assumptions on future demands and on the VLT. The model is trained to capture the behavior of optimal solutions from a perfect knowledge of the future. Collaborated with an industrial partner, our proposed E2E model has been implemented in production and we conduct a series of numerical experiments including a field experiment to demonstrate the advantage of the proposed E2E model over conventional two-step PTO approaches and current practices in industry. Our model, as well as the "E2E" philosophy, can be practically useful for the industry because it shortens the decision process and provides a more automatic inventory management solution. With the possibility of scaling and generalization, the proposed E2E model enables higher inventory-management accuracy with a lower operational cost and fewer labor efforts.

Moreover, we suggest several opportunities for future research in the E2E concept for supply-chain management. For instance, one potential direction would be trying to generalize the E2E model to more general inventory management settings, such as multi-echelon cases and inventory allocation problems. Another appealing direction would be constructing an E2E solution for jointly deciding order quantity and ordering time.

Chapter 3

Integrated Conditional Estimation-Optimization

3.1 Background and Motivation

Two fundamental aspects of decision-making under uncertainty are estimation and optimization. Classically these two aspects are treated separately, with statistical and/or machine learning methodologies used to estimate the distributions of uncertain parameters based on data, resulting in a stochastic optimization problem to be solved for making a decision. In recent years, researchers and practitioners have increasingly recognized the significance of considering estimation and optimization in tandem ([18, 72, 39, 44]). Another salient feature of modern decision-making under uncertainty is the presence of *contextual* information, usually in the form of features/covariates, that can be leveraged to improve the estimation of the uncertain parameters. For example, contextual information such as temporal information, the presence of promotions, and economic indicators can be leveraged to refine the estimation of uncertain demand for products. The refined demand distribution estimates would then be used for making inventory and supply chain decisions through optimization models. *Contextual stochastic optimization (CSO)* has recently emerged as a general paradigm describing this situation, with applications in supply chain management, finance, transportation, energy systems, and many other areas.

In this work, we consider the CSO problem in a data-driven setting where one has available historical data consisting of realizations of the uncertain parameters paired with contextual feature information. As mentioned, the classical method of solving CSO given data is a two-step procedure, where in the first step either a point prediction of the parameter or an estimation of its distribution is built based on data. (Although the phrases "prediction" and "estimation" are often synonymous or not clearly distinguished in the literature, herein we specifically let "prediction" denote point predictions of the random parameter and let "estimation" refer to any methodology, either parametric or non-parametric, for estimating

the conditional distribution of the random parameter given the context.) Modern machine learning techniques are often utilized in the first step to provide more granular results, and these models are usually fit based on statistical objectives such as measures of prediction error or likelihood. Then in the second step, given the prediction or estimation, an optimization problem is solved. A major drawback of these standard predict-then-optimize (PTO) and estimate-then-optimize (ETO) approaches is that they do not consider the decision error – the cost with respect to the downstream optimization problem due to an imperfect prediction – when fitting a statistical model.

We propose an integrated conditional estimation-optimization (ICEO) approach that estimates the underlying conditional distribution of the random parameter based on minimizing the ultimate decision error. We propose a highly flexible framework that models the conditional distribution using a hypothesis class and apply ideas from statistical learning to do estimation. As compared to existing approaches, our approach uses a generic learning framework based on specifying a hypothesis class and applies to a broad class of convex contextual stochastic optimization problems with uncertainty in the objective. Many previous approaches either rely heavily on the structure of the downstream problem, for example a linear ([44]) or newsvendor ([10]) problem, or propose a variation on a specific learning algorithm like random forests ([70]). In addition, related approaches based on "end-to-end learning" (see, e.g., [39, 150]) usually do not directly model the conditional distribution of uncertain parameters as we do in the ICEO framework and lack strong theoretical guarantees. We further discuss the relationship between the ICEO framework and existing approaches in Section 3.1.1.

In addition to proposing the flexible ICEO framework that models the conditional distribution of uncertain parameters in a way that accounts for the downstream optimization cost, we consider the statistical and computational properties of our approach. In particular, we prove asymptotic consistency in terms of risks, decisions and hypotheses. Asymptotic consistency is highly desired for data-driven methods because it guarantees that, as the amount of data increases, our solutions and estimated models converge to their optimums given full information of the true distribution of contextual features and uncertain parameters. We prove asymptotic consistence of the ICEO risk and induced decisions only under the assumption that the hypothesis class is compact, and consistency of the hypothesis requires an additional assumption related to the uniqueness of the true hypothesis. We also provide generalization bounds to quantify the out-of-sample performance when data is limited to a finite sample. To induce desirable generalization properties, we introduce a strongly convex decision regularization function to stabilize the ICEO decision and to eliminate potential multiple optimal decisions. The strongly convex regularization guarantees the Lipschitz property of the regularized optimal solution mapping, and the resulting generalization bounds are constructed based on multi-variate Rademacher complexity. In terms of computation, the core training problem of the ICEO framework is non-convex and even non-differentiable in many cases. In fact, due to the presence of constraints in the downstream problem, it is often

the case that the optimal decision oracle has a piece-wise constant shape, which leads to poor local minima that are very hard to escape when applying gradient-based methods. For these reasons, we propose two computational approaches: *(i)* a highly practical approach that involves approximating the regularized optimal solution oracle with a smooth function and then applying gradient algorithms, and *(ii)* a polynomial optimization approach when the downstream problem has a semi-algebraic objective and we approximate the optimal solution oracle with a polynomial function.

Our key contributions are summarized as follows:

1. We propose the ICEO framework, wherein we directly estimate the underlying conditional distribution of uncertain parameters given contextual information using a hypothesis class. In contrast to two-step ETO methods, we learn the conditional distribution in a way that integrates with the downstream optimization goal. ICEO offers more flexibility compared to most existing related approaches.

2. We prove asymptotic consistency of the ICEO method when the model i specified correctly (Theorem 3.3.1). More specifically, we show the consistency of ICEO risk, ICEO decisions, and ICEO hypothesis when the hypothesis class contains the correct conditional distribution function. The consistency in risk and decisions hold for arbitrary compact hypothesis classes. To guarantee the consistency in hypothesis, we require an additional assumption related to the uniqueness of the true hypothesis.

3. To quantify the out-of-sample performance with finite samples, we provide generalization bounds for the ICEO method based on the multi-variate Rademacher complexity of the hypothesis class used to learn the conditional distribution (Theorem 3.3.3). The generalization bound is based on the Lipschitz property of the regularized optimal decision oracle (Proposition 3.3.1.1).

4. The ICEO training problem is non-convex and non-differentiable. Non-differentiability poses a serious concern when applying gradient-based algorithms, like (stochastic) gradient descent, to solve the ICEO training problem as the presence of constraints can lead to local minima that are hard to escape (visually illustrated in Figure 3.1). To address this issue, we approximate the oracle using differentiable function classes with a guaranteed approximation error (Proposition 3.4.1.1, Proposition 3.4.1.2). We then provide corresponding generalization bounds when training ICEO method using the approximated oracle (Theorem 3.4.1). In addition, for the case where the nominal optimization problem is semi-algebraic, we propose an exact solution algorithm (Proposition 3.4.1.3).

The remainder of this chapter is organized as follows. In Section 3.1.1, we review related methods in literature. The details of our proposed ICEO framework are introduced in

Section 3.2. In Section 3.3, we provide performance guarantees in terms of asymptotic consistency and generalization bounds. In Section 3.4, we discuss the main difficulties in solving the ICEO formulation and provide solution methods. Empirical performance of the ICEO method is demonstrated in Section 3.5.

3.1.1 Relevant Literature

The fusion of prediction models based on data and the optimization problems has become more and more widespread in recent years. In the remainder of this section, we will discuss existing works related to this topic and contrast them with our proposed ICEO approach.

The first stream of research focuses on providing a prescriptive solution by approximating the conditional distribution of the random parameter given a feature vector, with the help of various machine learning tools. [18] first proposed prescriptive models that approximate the conditional distribution with a weighted empirical distribution of the uncertainty. The weights can be achieved based on multiple machine learning models, including k-nearest neighbors (KNN), kernel methods, tree-based methods, etc. A later work [19] investigates these prescriptive methods in the multi-period problem setting. [16] follows the same idea and propose a tree-based algorithm that balances the optimality of the prescription and accuracy of the prediction. [70] consider a random forest model for the prescriptive solution. In contrast to the standard way of splitting the feature space, the authors consider the down stream optimization quality while constructing the partitions. [61] considers the regularized Nadaraya-Watson approach and establish performance guarantees using moderate deviations theory.

Another stream of related work investigates adjusting the loss function to meet the ultimate optimization goal while training the machine learning models to predict the random parameters. [10] investigates the Newsvendor problem, which is inherently equivalent to a quantile prediction problem. The authors learn the feature-to-decision mapping from data by adopting a loss function that characterizes the newsvendor inventory cost, and is equivalent to the quantile loss function. Following a similar setting, [116] discuss the performance guarantees of such an approach when there are inter-temporal dependencies and non-stationarities. [44] consider the case when the downstream optimization problem has a linear objective. The authors propose a "smart predict-then-optimize" (SPO) framework with a tractable convex surrogate loss function (SPO+) to integrate the ultimate optimization problem structure. They prove Fisher consistency of SPO+ and demonstrate its strong numerical performance on different problem classes. [9] later provide finite-sample performance guarantee of the SPO loss in the form of generalization bounds. Recently, [95] have strengthened the consistency of SPO+ by providing risk guarantees and a calibration analysis in the polyhedral and strongly convex cases. [43] propose a method to train decision trees using the SPO loss and demonstrate strong numerical performance and improved model complexity over standard decision tree methods (e.g., CART) that minimize prediction error.

Other existing studies aim to learn the task-based end-to-end learning models with differentiable optimization layers. [39] consider a general setting where the optimization stage involves a convex optimization problem and adopt the objective in the optimization stage as the loss function to achieve an end-to-end training for the machine learning models. The main issue in such end-to-end learning models is to address the non-differentiability of the optimal solution mapping (the mapping from a contextual feature vector to the optimal decision). [7] introduce the differentiable optimization layers for the end-to-end training approaches and propose a method of approximating the gradient of the optimal solution mapping by the solution of a group of equations representing the KKT conditions. [2] further provide a method to convert convex programs to the canonical forms that can be implemented at the optimization layer and implemented their grammar in CVXPY for ease of use. [149] and [150] further consider more difficult combinatorial problems. They propose end-to-end models that map from the graph structure to a feasible solution and train them with the quality of the solution. [149] consider continuous relaxations of the discrete problem to propagate gradients through the optimization procedure. [97] consider mixed integer linear programs and consider a homogeneous self-dual formulation of the LP and show that the gradients are related to an interior point step. [15] instead consider stochastically perturbed optimizers to evaluate the gradients required for back-propagation. [98, 49, 113] also discuss how to approximate the gradients when training end-to-end models for combinatorial problems. As our work focuses on convex optimization problems, we skip the details and refer to [85] for a detailed survey. Although demonstrated to be competitive in numerical experiments, these end-to-end learning models based on optimization layers and their extensions to combinatorial cases lack strong performance guarantees in theory. Moreover, learning the feature-to-decision mapping lacks flexibility in the way that it handles constraints. Indeed, constraints restricts the hypothesis class that can be used to learn the data-to-decision mapping. In contrast, our ICEO framework learns the conditional distribution and use the optimal solution mapping to obtain the decision, which is more flexible in handling constraints.

We also comment on the difference between the problem setting of ICEO and the joint estimation-optimization (JEO) model ([68, 3, 67, 107]). The major difference is that, in the JEO model, there is no contextual information considered as predictors of the uncertainty. Besides, several JEO models focus on solving an online convex optimization problem in the optimization stage, while we consider a stochastic optimization problem. We would also like to point out the differences in the problem setting of ICEO and the operational statistics method, in which the downstream optimization goal is considered in finding the optimal operational statistic ([96, 36, 121]). We include the contextual information in our problem setting which is not considered in the classic operational statistics literature. Moreover, we aim to learn the underlying conditional distribution rather than finding the best statistic. We also consider constraints in the downstream optimization problem.

Other related works include [108], which investigates the relationship between the pre-

diction part to the performance of the optimization part, mainly in the case of the least squares loss function. [28] focuses on the mean-variance portfolio optimization problem and integrates regression based predictive models with the optimization setting. The authors provide closed-form analytical solutions for the unconstrained cases. [117] instead focuses on a multi-period inventory management problem with random demand and leadtime, and provide a practical end-to-end learning framework empowered by deep learning models. The authors demonstrate the empirical success of this approach in practice by conducting a field experiment in industry.

3.2 Contextual Stochastic Optimization and the ICEO Approach

In this section, we review the basic ingredients of contextual stochastic optimization problems, which is a fundamental model for applying machine learning in many operational contexts, and we formally describe our ICEO approach. We consider a convex CSO, which models a downstream decision-making task. The feasible region for the decision variable $w \in \mathbb{R}^d$, denoted by $S \subset \mathbb{R}^d$, is assumed to be known with certainty. We additionally assume that S is a convex and compact set. Although the feasible region of our optimization task is known with certainty, the objective function $c(\cdot, \xi) : S \to \mathbb{R}$ is stochastic and depends on a random parameter ξ. We assume that, for all values of ξ, $c(\cdot, \xi)$ is a convex function of w. While the precise value of ξ is not known at the time when a decision must be made, we assume that the decision maker observes an associated contextual feature vector $x \in \mathcal{X} \subseteq \mathbb{R}^p$ (sometimes the components of x are referred to as covariates) that can be used to learn information about the objective function. Let $\mathcal{D}$ denote the joint distribution of x and ξ. Then, given an observed $x \in \mathbb{R}^p$, the decision maker's goal is to solve the contextual stochastic optimization problem:

$$\min_{w \in S} \ \mathbb{E}_\xi[c(w, \xi)|x], \tag{3.1}$$

where the expectation above is with respect to the *conditional distribution* of ξ given x.

It is important to emphasize that the distribution $\mathcal{D}$, and hence the conditional distribution of ξ given any x, is typically unavailable in practice. Instead, a data-driven approach to solving (3.1) is much more viable. Indeed, one often has available a training dataset $\{(x_i, \xi_i)\}_{i=1}^n$ consisting of historically observed pairs of feature vectors $x_i \in \mathcal{X}$ and associated parameter values ξ_i. If the dataset $\{(x_i, \xi_i)\}_{i=1}^n$ is an independent and identically distributed sample from the distribution $\mathcal{D}$, for example, then it may be possible to learn enough information about the conditional distribution to solve problem (3.1). Note also that, as pointed out by [18] for example, due to the nonlinearity of the objective function, a point estimate for a prediction of ξ given x usually does not provide enough information about the conditional distribution to produce a reasonable solution of (3.1). In general, without any additional

structural assumptions, e.g., on either the random parameter ξ, the distribution $\mathcal{D}$, or the cost functions $c(\cdot,\cdot)$, adequately learning the conditional distribution for all relevant x may be an intractable problem.

In this work, we consider the case where the random parameter ξ has finite discrete support, i.e., $\xi \in \Xi := \{\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_K\}$. Then, for any $x \in \mathcal{X}$, the conditional distribution of ξ given x is characterized by a probability vector $p^*(x) \in \Delta_K$, where $\Delta_K := \{p \in \mathbb{R}^K : \sum_{k=1}^K p_k = 1, p \geq 0\}$ denotes the $(K-1)$-dimensional unit simplex. That is, $p_k^*(x)$, the k-th component of $p^*(x)$, is defined by $p_k^*(x) = \mathbb{P}_\xi(\xi = \tilde{z}_k|x)$, for all $k = 1, \ldots, K$. Using this notation as well as the shorthand notation $c_k(\cdot) := c(\cdot, \tilde{z}_k)$ for all $k = 1, \ldots, K$, problem (3.1) can be equivalently written as

$$\min_{w \in S} \mathbb{E}_\xi[c(w,\xi)|x] \;=\; \min_{w \in S} \sum_{k=1}^K p_k^*(x) c_k(w). \tag{3.2}$$

3.2.1 The ICEO Approach

Let us now describe the major ingredients of our ICEO approach, as well as the formulation of our ICEO training problem.

3.2.1.0.1 Hypothesis Class of Conditional Probability Estimators It is evident from the right side of (3.2) that learning the conditional distribution $p^*(x)$ is the most critical part of our contextual stochastic optimization setting. We adopt standard ideas from learning theory to learn $p^*(x)$, whereby we employ a compact hypothesis class $\mathcal{H}$ of conditional probability estimators. That is, $\mathcal{H}$ is a compact set (e.g., with respect to the uniform norm) of functions $f : \mathcal{X} \to \Delta_K$. The hypothesis class $\mathcal{H}$ is the first major ingredient of our ICEO approach. Note that the constraint on the output of $f \in \mathcal{H}$, namely $f(x) \in \Delta_K$ for all $x \in \mathcal{X}$, is not standard in most learning problems but is necessitated by our setting. Fortunately, this constraint can be accommodated in a number of ways. A straightforward approach is to consider the softmax operator soft : $\mathbb{R}^K \to \mathbb{R}^K$ defined by $\text{soft}_k(v) = \frac{\exp(v_k)}{\sum_{j=1}^K \exp(v_j)}$ for $v \in \mathbb{R}^K$. Then, given *any* hypothesis class $\tilde{\mathcal{H}}$ of unconstrained functions $\tilde{f} : \mathcal{X} \to \mathbb{R}^K$, we can define $\mathcal{H}$ as the composition class $\text{soft} \circ \tilde{\mathcal{H}}$. Note that, due to the differentiability properties of the softmax operator, $\text{soft} \circ \tilde{\mathcal{H}}$ naturally inherits differentiability properties from $\tilde{\mathcal{H}}$, which can be very useful from a computational perspective. For another example, consider $\mathcal{H}$ defined by a decision tree partitioning algorithm. Then, for any given x, $f(x)$ can be constructed from the empirical distribution of ξ restricted to the subset of the partition of the training data for which x lies in. Finally, a third approach, which we expand upon in Section 3.4.3, is to let $\mathcal{H}$ be a constrained linear hypothesis class whereby $\mathcal{H} = \{f : f(x) = Bx \in \Delta_K \text{ for all } x \in \mathcal{X}\}$. Depending on the structure of $\mathcal{X}$, it may be possible to efficiently model the constraint $Bx \in \Delta_K$ for all $x \in \mathcal{X}$, and we discuss specific examples in Section 3.4.3. We would like to emphasize two points about our approach for estimating

the conditional distribution using a hypothesis class $\mathcal{H}$. First, by directly estimating the conditional probability our proposed method has more flexibility in handling constraints as compared to methods that learn a mapping π directly from features x to decisions w. In particular, the approach of learning a mapping from features to decisions requires that the output of the mapping π be feasible in the region S, which may severely constrain the feasible set of π. On the other hand, our approach of composing a user-specified hypothesis class $\mathcal{H}$ with the regularized optimal solution mapping $w_\rho(\cdot)$ allows for a very general selection of $\mathcal{H}$. In particular, the only requirement to achieve asymptotic consistency is compactness of $\mathcal{H}$, and, to provide a generalization bound, we further need $\mathcal{H}$ to have bounded multivariate Rademacher complexity.

3.2.1.0.2 Regularized Optimization Oracle. As mentioned previously, we assume that the functions $c_k(\cdot) = c(\cdot, \tilde{z}_k)$, for all $k = 1, \ldots, K$, are all convex functions of w on the convex and compact feasible region S. We additionally assume that these functions are computationally tractable in practice, in the sense that any weighted combination of these functions can be efficiently optimized. Furthermore, we presume that we can additionally work with a *decision regularization function* $\phi(\cdot) : S \to \mathbb{R}$, which is non-negative and strongly convex with respect to some norm $\|\cdot\|$ on $\mathbb{R}^d$. Given any $p \in \Delta_K$ and $\rho > 0$, define the regularized optimal solution mapping:

$$w_\rho(p) := \arg\min_{w \in S} \sum_{k=1}^{K} p_k c_k(w) + \rho\phi(w). \tag{3.3}$$

Note that, due to the strong convexity of $\phi(\cdot)$, $w_\rho(p)$ is uniquely defined. Furthermore, we can show that $w_\rho(\cdot)$ is a continuous mapping as demonstrated in Lemma B.1.1. These regularity properties induced by the use of the regularization term $\phi(\cdot)$ are crucial for developing our ICEO methodology as well as for proving associated theoretical guarantees. For computational purposes, we assume that $w_\rho(p)$ can be efficiently computed in practice for any $p \in \Delta_K$ and $\rho > 0$. For example, we may compute $w_\rho(p)$ using a commercial solver or a specialized algorithm that depends on the structure of the $c_k(\cdot)$ and $\phi(\cdot)$ functions. Ideally, the function $\phi(\cdot)$ should be chosen so that the complexity of computing $w_\rho(p)$ is not greatly increased as compared to when $\rho = 0$. Note that our performance guarantees developed in Section 3.3 hold for any choice of $\phi(\cdot)$ that is strongly convex. When $\phi(\cdot)$ is not present there may be multiple optimal solutions of (3.3), and we use the notation $W(p)$ to refer to the set of such optimal solutions, i.e., $W(p) := \arg\min_{w \in S} \sum_{k=1}^{K} p_k c_k(w)$.

3.2.1.0.3 ICEO Methodology. We are now ready to describe our ICEO methodology and corresponding training problem, whereby we consider an integrated approach that estimates a hypothesis $f \in \mathcal{H}$ in consideration of the downstream optimization goal. We presume that we have collected a training dataset $\{(x_i, \xi_i)\}_{i=1}^n$ consisting of historically observed pairs of feature vectors $x_i \in \mathcal{X}$ and associated parameter values ξ_i. We also presume that the decision maker uses the regularized optimal solution oracle $w_\rho(\cdot)$ defined in (3.3).

We adopt the empirical risk minimization (ERM) principle with respect to the regularized in-sample cost induced by the regularized oracle:

$$\begin{aligned} \min_{f\in\mathcal{H}, w_1,\ldots,w_n\in S} \quad & \frac{1}{n}\sum_{i=1}^{n} c(w_i,\xi_i) + \rho\phi(w_i) \\ \text{s.t.} \quad & w_i = w_\rho(f(x_i)), \end{aligned} \tag{ICEO-ρ}$$

where $\rho > 0$ is a given value of the decision regularization parameter, which can be chosen with cross validation for example. Let $\hat{f} \in \mathcal{H}$ denote a computed optimal solution of (ICEO-ρ). Then, for any newly observed feature vector $x \in \mathcal{X}$, the decision maker implements the decision $w_\rho(\hat{f}(x)) \in S$ formed by composing $w_\rho(\cdot)$ with $\hat{f}(\cdot)$.

Let us contrast the ICEO approach with two more standard approaches: predict-then-optimize (PTO) and estimate-then-optimize (ETO). Note that the phrases "predict" and "estimate" are closely tied in the literature and there is no agreed upon consistent way to distinguish between the two. In our context, we specifically use "predict" to refer to point predictions of the random parameter ξ and "estimate" to refer to any methodology for estimating the conditional distribution of ξ given any $x \in \mathcal{X}$. Specifically, in the PTO approach, a machine learning model $\hat{g}_{\text{PTO}} : \mathcal{X} \to \Xi$ is built, using the training data, to predict the parameter ξ based on the feature vector x. Then, given any new $x \in \mathcal{X}$, the decision maker implements a decision from the optimal solution set $\arg\min_{w\in S} c(w, \hat{g}_{\text{PTO}}(x))$. As mentioned previously, due to the nonlinearity of the objective, a point estimate for a prediction of ξ is generally too simplistic to provide a reasonable solution of (3.1). Indeed, unless the conditional distribution is guaranteed to be a point mass, then the PTO approach is not suitable for nonlinear problems. In the case when the objective is linear, a point estimate is actually sufficient and PTO approach is viable. In this linear case, [44] propose a "smart predict-then-optimize (SPO)" approach that aims to minimize the downstream optimization cost. Furthermore, in this linear case the ICEO approach proposed herein (without regularization) reduces to the SPO problem proposed by [44].

Returning to the nonlinear case studied herein, the ETO approach learns a model $\hat{f}_{\text{ETO}} : \mathcal{X} \to \Delta_K$ for estimating the conditional distribution of ξ given x. Then, given any new $x \in \mathcal{X}$, the decision maker implements a decision from the optimal solution set $W(\hat{f}_{\text{ETO}}(x))$. Thus, the ETO approach is more aligned with the ICEO approach. The main distinction is that the traditional ETO approach learns the model $\hat{f}_{\text{ETO}}$ in a way that is completely oblivious to the downstream optimization task. For instance, given a hypothesis class $\mathcal{H}$, the ETO approach might select the hypothesis by minimizing the empirical cross-entropy loss, defined for any $f \in \mathcal{H}$ and any observed $(x, \xi = \tilde{z}_k)$ by $\ell_{\text{ce}}(f(x), \xi = \tilde{z}_k) := -\log(f_k(x))$. Alternatively, one may consider a purely non-parametric method for estimating the conditional distribution such as the k-nearest neighbors or CART algorithms for example. In these cases, and several others, [18] demonstrate asymptotic consistency properties of the ETO approach. [70] consider using a (non-parametric) random forests estimator of the conditional distribution in

a way that is trained with respect to the cost of the downstream optimization task, akin to the ICEO approach. On the other hand, the ICEO approach directly models the underlying conditional distribution using a hypothesis class $\mathcal{H}$. Thus, while [70] only provide asymptotic consistency results, we are able to prove both asymptotic consistency and generalization bounds for a wide variety of hypothesis classes.

Additional Notation. Due to the compactness of S, the cost function $c(\cdot,\cdot)$ is bounded and we define $\bar{c} := \sup_{w\in S,\xi\in\Xi} |c(w,\xi)|$. Because of the compactness of S, we can define diameters of S. We let $\text{diam}_j(S) := \sup_{u,v\in S} |u_j - v_j|$ to denote the coordinate-wise diameter of the feasible region S. We further let $\text{diam}(S) := \sum_{j=1}^{d} \text{diam}_j(S)$ denote the summation of the coordinate-wise diameter of all coordinates. Given a norm $\|\cdot\|$ defined on $\mathbb{R}^d$, the distance from a point $w \in \mathbb{R}^d$ to a set $W \subseteq \mathbb{R}^d$ is denoted by $\text{dist}(w, W) := \inf_{u\in W} \|w - u\|$. For a convex function $h(\cdot) : S \to \mathbb{R}$, we let $\partial h(w)$ denote the set of subgradients of $h(\cdot)$ at w. Let $\circ$ denote the composition of functions. For example, with $f : \mathcal{X} \to \Delta_K$ and $w : \Delta_K \to S$, then $w \circ f$ is the function from $\mathcal{X}$ to S with $(w \circ f)(x) := w(f(x))$ for all $x \in \mathcal{X}$. This function composition notation also extends naturally to function classes. For example, for a class $\mathcal{H}$ of functions $f : \mathcal{X} \to \Delta_K$, we let $w \circ \mathcal{H}$ denote the class of functions $\{w \circ f : f \in \mathcal{H}\}$. We denote the set of non-negative integers a $\mathbb{N}_0$ and let $\mathbb{N}_0^k$ denote the set of all k-dimensional vectors with each component is a non-negative integer. $\mathbb{1}$ denotes the K-dimensional vector with all coordinates taking the value of one. We let $\text{TV}(\mathcal{P}, \mathcal{Q})$ denote the total variation between two probability measures $\mathcal{P}$ and $\mathcal{Q}$ supported on the K-dimensional simplex Δ_K. $\text{TV}(\mathcal{P}, \mathcal{Q}) := \sum_{A\in\mathcal{B}} |\mathcal{P}(A) - \mathcal{Q}(A)|$ where $\mathcal{B}$ denote the class of Borel sets in Δ_K. In Section 3.4.1, we will use an equivalent expression of $\text{TV}(\mathcal{P}, \mathcal{Q}) = \frac{1}{2} \sup_{f:\Delta\to[-1,1]}(\int_{\Delta_K} f(p)d\mathcal{P}(p) - \int_{\Delta_K} f(p)d\mathcal{Q}(p))$.

3.2.2 Motivating Examples

In this section, we present a few motivating examples for the ICEO framework, some of which will be revisited in our numerical experiments in Section 3.5.

Example 3.2.1 (Multi-item Newsvendor)**.** The multi-item Newsvendor problem aims to find the optimal replenishment quantities for d different products. We let $\xi := (\xi_1, \ldots, \xi_d)$ denote the random demand of d products and let $w \in \mathbb{R}^d$ denote the associated order quantities. The demand values ξ might be related to contextual information such as promotions, holiday seasons, brand information, etc. The objective of this problem is the total inventory cost including the holding costs h_l and stockout costs b_l, which characterize the over-stock and under-stock, respectively. The objective cost can be formulated as

$$c(w,\xi) := \sum_{l=1}^{d} h_l(w_l - \xi_l)^+ + b_l(\xi_l - w_l)^+, \tag{3.4}$$

where the function $(\cdot)^+$ is defined as $\max\{\cdot, 0\}$. Moreover, we consider a budget capacity

constraint $C > 0$ on the total order quantities and formulate the feasible set as

$$S := \{w : \sum_{l=1}^{d} w_l \leq C, w \geq 0\}.$$

Example 3.2.2 (Risk-Averse Portfolio Optimization)**.** We consider the problem of finding an optimal risk-averse portfolio of d assets. We denote the random vector of asset returns by $\xi \in \mathbb{R}^d$, which may be associated with the contextual information such as economic indicators, news headlines, etc. The decision maker aims to find the best allocation of assets $w \in \mathbb{R}^d$ that optimizes a weighted combination of the expected return and variance of the portfolio. By introducing an auxiliary variable $w_0 \in \mathbb{R}$, we formulate the objective as

$$c(w, w_0, \xi) := \alpha \left(\sum_{l=1}^{d} w_l \xi_l - w_0 \right)^2 - \sum_{l=1}^{d} w_l \xi_l, \tag{3.5}$$

where $\alpha > 0$ is a trade off parameter. Note that the expectation of the first term in (3.5) is $\alpha \mathbb{E}_\xi \left[(\sum_{l=1}^{d} w_l \xi_l - w_0)^2 \right]$, which represents the variance of the investment return $\text{Var}(\sum_{l=1}^{d} w_l \xi_l)$ when w_0 is optimally selected as $w_0 = \mathbb{E}_\xi \left[\sum_{l=1}^{d} w_l \xi_l \right]$, while the second term is the return of the portfolio. Therefore, $\mathbb{E}_\xi[c(w, w_0, \xi)]$ trades off between minimizing the variance and maximizing the expected return of the portfolio. As is standard in the classical portfolio optimization problems, we constrain the portfolio decision in the simplex $\Delta_d = \{w \in \mathbb{R}^d : \sum_{l=1}^{d} w_l = 1, w \geq 0\}$ and we have

$$S := \{(w, w_0) : w \in \Delta_d, w_0 \geq 0, 0 \leq w_0 \leq \bar{\Xi}\},$$

where $\bar{\Xi} \geq 0$ is a known upper bound the maximum of the returns $\|\xi\|_\infty$.

Example 3.2.3 (Minimum Convex Cost Flow Problem)**.** Many applications such as urban traffic system and area transfers in communication networks can be formulated as a minimum convex cost flow problem (we refer to Chapter 14 of [4] for more details). In the minimum convex cost flow problem, the decision-maker aims to find the maximum flow that minimizes the associated cost on the edges. The cost is a convex function of flow and depends on a random parameter. Suppose we consider a directed graph with d edges and the random parameter $\xi \in \mathbb{R}^d$. In this problem, we consider the objective function

$$c(w, \xi) = \sum_{l=1}^{d} g_l(w_l, \xi_l)$$

where g_l is a convex function of w_l and g_l can be different for different coordinates. Similar to the standard network flow problem, we let the matrix A denotes the node-arc incidence matrix of the graph and restrict the flow on each edge in the region $[l, u]$. Therefore, we have the feasible region

$$S = \{w \in \mathbb{R}^d : Aw = 0, w \in [l, u]^d\}.$$

3.3 Performance Guarantees

In this section, we demonstrate asymptotic consistency and finite-sample performance guarantees of the ICEO approach. Let us first introduce some additional notation. We state our results in terms of arbitrary policy mappings $\pi : \mathcal{X} \to S$, which represent any mapping from the feature space $\mathcal{X}$ to the set of feasible decisions S. Our main interest herein is the class of policies that combine the optimal solution mapping and hypothesis f, i.e., $\Pi = w_\rho \circ \mathcal{H}$. This class of policies includes the policy learned by the ICEO approach as well as policies learned by ETO approaches. In the remaining part of this work, we let $f^* : \mathcal{X} \to \Delta_K$ denote the function that maps from x to the true conditional distribution $p^*(x)$. We refer to f^* as the true hypothesis. Moreover, we define $w(\cdot) : \Delta_K \to S$ as a function that arbitrarily outputs a value from the optimal solution set $W(\cdot)$, i.e., $w(p) \in W(p) = \arg\min_{w \in S} \sum_{k=1}^K p_k c_k(w)$ for all $p \in \Delta_K$.

To quantify our performance guarantees, we define the following risk functions for any policy π and given regularization parameter $\rho \geq 0$:

1. $\hat{R}_n(\pi; \rho)$: The empirical regularized risk, for any given regularization parameter $\rho \geq 0$, with respect to a given sample $\{(x_i, \xi_i)\}_{i=1}^n$, i.e.,

$$\hat{R}_n(\pi; \rho) := \frac{1}{n} \sum_{i=1}^n c(\pi(x_i), \xi_i) + \rho\phi(\pi(x_i)).$$

2. $R(\pi; \rho)$: The expected regularized risk, for any given regularization parameter $\rho \geq 0$, with respect to the underlying joint distribution $\mathcal{D}$ of x and ξ, i.e.,

$$R(\pi; \rho) := \mathbb{E}_{x,\xi} \left[c(\pi(x), \xi) + \rho\phi(\pi(x)) \right] = \mathbb{E}_x \left[\sum_{k=1}^K p_k^*(x) c_k(\pi(x)) + \rho\phi(\pi(x)) \right],$$

   where $p_k^*(x) = \mathbb{P}_\xi(\xi = \tilde{z}_k | x)$ for all $k = 1, \ldots, K$.

We also use the short hand notation $\hat{R}_n(\pi) := \hat{R}_n(\pi; 0)$ and $R(\pi) := R(\pi; 0)$ to denote the unregularized empirical and expected risks, respectively. Note that $\hat{R}_n(\cdot; \rho)$ is the objective function of (ICEO-ρ). We also use the notation $\mathcal{D}_x$ to refer to the marginal distribution of the features x. We further define the optimal risk values for the class of policies $\Pi = w_\rho \circ \mathcal{H}$ that we consider herein.

1. J^*: the optimal expected unregularized risk, i.e.,

$$J^* := \min_{f \in \mathcal{H}} \mathbb{E}_x \left[\sum_{k=1}^K p_k^*(x) c_k(w(f(x))) \right] = \min_{f \in \mathcal{H}} R(w \circ f; 0).$$

2. J_ρ^*: the optimal expected regularized risk for any given regularization parameter $\rho > 0$, i.e.,

$$J_\rho^* := \min_{f\in\mathcal{H}} \mathbb{E}_x \left[\sum_{k=1}^{K} p_k^*(x) c_k(w_\rho(f(x))) + \rho\phi(w_\rho(f(x)))\right] = \min_{f\in\mathcal{H}} R(w_\rho \circ f; \rho).$$

3. $\hat{J}_\rho^n$: the optimal empirical regularized risk with any given sample S_n and a given regularization parameter $\rho > 0$, i.e.,

$$\hat{J}_\rho^n := \min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} c(w_\rho(f(x_i)), \xi_i) + \rho\phi(w_\rho(f(x_i))) = \min_{f\in\mathcal{H}} \hat{R}_n(w_\rho \circ f; \rho),$$

and we let $\hat{f}_\rho^n$ denote its optimal solution.

3.3.1 Asymptotic Consistency

We first demonstrate the asymptotic consistency of the ICEO approach. The consistency of our approach is three-fold: the consistency of the ICEO risk, the consistency of the ICEO decisions, and the consistency of the ICEO hypothesis. Our asymptotic consistency results require the following conditions:

Assumption 1. *For the compact hypothesis class $\mathcal{H}$, we have the following:*

A. (Model Specification) The hypothesis class $\mathcal{H}$ includes the true hypothesis f^* i.e., $f^* \in \mathcal{H}$.

B. (Unique true hypothesis.) There does not exists a hypothesis $f \neq f^*$ in $\mathcal{H}$ such that $W(f(x)) \cap W(f^*(x)) \neq \emptyset$, $\mathcal{D}_x$-almost surely for all $x \in \mathcal{X}$.

Recall that if $c(\cdot)$ is not strongly convex, there may exist multiple optimal solutions denoted by $W(p)$ for any input probability vector $p \in \Delta_K$. In this case, we assume that the oracle $w(p)$ outputs one element from the set of optimal solutions $W(p)$. In Lemma B.1.2 in the Appendix, we demonstrate that the value of the unregularized optimal riskJ^* does not depend on the particular choice of $w(\cdot)$ and that f^* is always the minimizer that achieves J^*.

To guarantee the consistency of the ICEO method, we consider a sequence of regularization parameters ρ_n, depending on the sample size n, such that ρ_n converges to zero as n grows to infinity. Theorem 3.3.1 below demonstrates the three levels of consistency.

Theorem 3.3.1 (Asymptotic Consistency of ICEO)**.** *Suppose that the training data (x_i, ξ_i) is an i.i.d. sequence from the distribution $\mathcal{D}$ and that the sequence of regularization parameters ρ_n satisfies $\lim_{n\to\infty} \rho_n = 0$. Then, under Assumption 1.A, we have the following:*

(i) The optimal empirical regularized risk converges to the optimal expected risk, i.e., $\hat{J}^n_{\rho_n} \to J^*$ *with probability 1.*

(ii) $\mathcal{D}_x$*-almost surely for all* $x \in \mathcal{X}$*, the sequence of ICEO decisions* $w_{\rho_n}(\hat{f}^n_{\rho_n}(x))$ *converges to the true set of optimal decisions* $W(f^*(x))$*, i.e.,* $\text{dist}(w_{\rho_n}(\hat{f}^n_{\rho_n}(x)), W(f^*(x))) \to 0$ *with probability 1.*

(iii) Additionally, with Assumption 1.B, the sequence of ICEO hypotheses converges to the true hypothesis, i.e, $\hat{f}^n_{\rho_n} \to f^*$ *with probability 1.*

We would like to clarify the relationship between the asymptotic consistency stated in Theorem 3.3.1 and the asymptotic optimality defined in [18]. In [18], the authors define asymptotic optimality in terms of the decisions reaching the best objective function performance possible. Because of the continuity of the cost function c, the convergence of ICEO decisions, as stated in (ii) of Theorem 3.3.1, implies the asymptotic optimality property of [18].

Proof. Proof of Theorem 3.3.1 In this proof, we slightly abuse the notation and let $R(f;\rho)$ denote $R(w_\rho \circ f;\rho)$ for any $\rho \geq 0$ and $\hat{R}_n(f;\rho)$ denote $\hat{R}_n(w_\rho \circ f;\rho)$. We first show that $\lim_{\rho\to 0} J^*_\rho = J^*$. Recall that $J^* = R(f^*;0)$ and $J^*_\rho = R(f^*_\rho;\rho)$, and we have:

$$R(f^*;0) \leq R(f^*_\rho;0) \leq R(f^*_\rho;\rho) \leq R(f^*;\rho),$$

where the first inequality holds because the true hypothesis f^* achieves the optimal value J^*, the second follows from the fact that ϕ is a non-negative function, and the third follows from the fact that f^*_ρ is the optimizer of $R(\cdot;\rho)$. In the meanwhile, $R(f^*;\rho) \to R(f^*;0)$ as $\rho \to 0$. Thus,

$$J^*_\rho := R(f^*_\rho;\rho) \to R(f^*;0) = J^*.$$

Then we want to show that $\hat{J}^n_\rho \to J^*_\rho$ with probability 1 as $n \to \infty$. Let $l_i(f,\rho) := c(w_\rho(f(x_i)),\xi_i) + \rho\phi(w_\rho(f(x_i)))$, then $l_i(\cdot,\rho)$ are i.i.d. random functions on the compact hypothesis class $\mathcal{H}$. Due to the compactness of S, both $R(\cdot;\rho)$ and $l_i(\cdot,\rho)$ are bounded. Then we can apply the main theorem in [126] and have that $\frac{1}{n}\sum_{i=1}^n l_i(\cdot,\rho) \to R(\cdot;\rho)$ with probability 1 for all f uniformly. Moreover, together with the continuity of $R(\cdot;\rho)$, the uniform convergence of $\hat{R}_n$ leads to the Γ-convergence of $\hat{l}(\cdot,\rho)$ to $R(\cdot;\rho)$ in probability ([26]). Then we can apply the Fundamental Theorem of Γ-convergence so that the convergence of minimum values $\hat{J}^n_\rho$ converges to J^*_ρ ([27]). Then for any sequence of $\rho_n > 0$ with $\rho_n \to 0$, as $n \to \infty$, we can find a sequence of $\epsilon_n > 0$ that satisfies $\lim_{n\to\infty} \epsilon_n = 0$ and the following conditions: $J^*_{\rho_n} - J^* \leq \epsilon_n$ and $|\hat{J}^n_{\rho_n} - J^*_{\rho_n}| < \epsilon_n$ with probability 1. Thus, $|\hat{J}^n_{\rho_n} - J^*| \leq 2\epsilon_n$ for all n, which leads to $\hat{J}^n_{\rho_n} \to J^*$ with probability 1 and *(i)* is proved.

Due to the compactness of S and $\mathcal{H}$, with any sequence of $\rho_i \to 0$, the sequence $w_{\rho_i}(f^*_{\rho_i}(x))$ has accumulation points. Let $w_{\rho_t}(f^*_{\rho_t}(x))$ be any subsequence converging to an accumulation point $w_{\rho_\infty}(f^*_{\rho_\infty}(x)) \in S$. Note that we have

$$\begin{aligned} J^* &= \lim_{t\to\infty} \mathbb{E}_x[\sum_{k=1}^K f_k^*(x)c_k(w_{\rho_t}(f^*_{\rho_t}(x))) + \rho_t\phi(w_{\rho_t}(f^*_{\rho_t}(x)))] \\ &\geq \lim_{t\to\infty} \mathbb{E}_x[\sum_{k=1}^K f_k^*(x)c_k(w_{\rho_t}(f^*_{\rho_t}(x)))] \\ &\geq \mathbb{E}_x[\sum_{k=1}^K f_k^*(x)c_k(w_{\rho_\infty}(f^*_{\rho_\infty}(x)))], \end{aligned}$$

where the equality follows from $J^*_\rho \to J^*$ when $\rho \to 0$, the first inequality holds due to the non-negativity of the regularization term ϕ, and the second by Fatou's lemma. Since $W(f^*(x))$ is defined the set of optimal solutions of $\sum_{k=1}^K f_k^*(x)c_k(\cdot)$ for all x, then $\mathcal{D}_x$ almost surely for all x, $w_{\rho_\infty}(f^*_{\rho_\infty}(x))$ must lie in the set $w(f^*(x))$. Then $\text{dist}(w_{\rho_t}(f^*_{\rho_t}(x)), W(f^*(x))) \leq \|w_{\rho_t}(f^*_{\rho_t}(x)) - w_{\rho_\infty}(f^*_{\rho_\infty}(x))\|$. By the continuity of the norm $\|\cdot\|$, $\lim_{t\to\infty}\|w_{\rho_t}(f^*_{\rho_t}(x)) - w_{\rho_\infty}(f^*_{\rho_\infty}(x))\| = \|\lim_{t\to\infty} w_{\rho_t}(f^*_{\rho_t}(x)) - w_{\rho_\infty}(f^*_{\rho_\infty}(x))\| = 0$. Therefore, $\mathcal{D}_x$-almost surely for all x, $\text{dist}(w_\rho(\hat{f}^n_\rho(x)), W(f^*(x))) \to 0$ as $\rho \to 0$.

Moreover, for any fixed $\rho > 0$, due to the compactness of $\mathcal{H}$ and the fact that $\frac{1}{n}\sum_{i=1}^n l_i(\cdot,\rho)$ converges to $R(\cdot;\rho)$ uniformly with probability 1, we can apply the fundamental theorem of Γ-convergence again and conclude that any accumulation point of $\hat{f}^n_\rho$, denoted by f^∞_ρ, minimizes $R(\cdot;\rho)$ with probability 1. Because of the strongly convexity of $c_k(\cdot) + \rho\phi(\cdot)$, if f^∞_ρ minimizes $R(\cdot;\rho)$, then $w(f^\infty_\rho(x)) = w(f^*_\rho(x))$ $\mathcal{D}_x$-almost surely for all $x \in \mathcal{X}$. Therefore, given any sequence $\rho_n \to 0$, we can find a sequence of $\delta_n \geq 0$ such that $\lim_{n\to\infty}\delta_n = 0$ and satisfies both $\|w(f^*_{\rho_n}(x)) - w_{\rho_n}(\hat{f}^n_{\rho_n}(x))\| \leq \delta_n$ and $\text{dist}(w_{\rho_n}(f^*_{\rho_n}(x)), W(f^*(x))) \leq \delta_n$, with probability 1. Therefore, with probability 1, we have

$$\begin{aligned} \text{dist}(w_{\rho_n}(\hat{f}^n_{\rho_n}(x)), W(f^*(x)) &= \inf_{u\in W(f^*(x))} \|w_{\rho_n}(\hat{f}^n_{\rho_n}(x)) - w_{\rho_n}(f^*_{\rho_n}(x)) + w_{\rho_n}(f^*_{\rho_n}(x)) - u\| \\ &\leq \|w_{\rho_n}(\hat{f}^n_{\rho_n}(x)) - w_{\rho_n}(f^*_{\rho_n}(x))\| + \inf_{u\in W(f^*(x))} \|w_{\rho_n}(f^*_{\rho_n}(x)) - u\| \qquad (3.6) \\ &\leq 2\delta_n, \qquad \mathcal{D}_x - \text{almost surely}, \qquad (3.7) \end{aligned}$$

where (3.6) follows from the triangle inequality, and (3.7) further leads to the conclusion in *(ii)*.

Finally, if we have an accumulation point of $\hat{f}^n_{\rho_n}$, denoted as $\bar{f}$, that does not equal to f^*, then by *(ii)*, we have $W(\bar{f}(x)) \cap W(f^*(x)) \neq \emptyset$ for all $x \in \mathcal{X}$ almost surely, which contradict to Assumption 1.B. Thus with the uniqueness assumption, the true hypothesis f^* can be recovered by $\hat{f}^n_{\rho_n}$. □

3.3.2 Finite Sample Performance Guarantees

We now provide finite sample performance guarantees of the ICEO solution $\hat{f}^n_{\rho_n}$ in the form of generalization bounds based on Rademacher complexities. In particular, our overall strategy is as follows: *(i)* we demonstrate that, due to the presence of the strongly convex decision regularization function $\phi(\cdot)$, the optimal solution mapping $w_\rho(\cdot)$ is Lipschitz, *(ii)* we use the result of [99] to bound the Rademacher complexity with respect to the cost function of the ICEO framework by the multivariate Rademacher complexity of the underlying hypothesis class $\mathcal{H}$. In addition, we slightly abuse the notation and let $c(\cdot) : S \to \mathbb{R}^K$ denote a vector-valued mapping, where each component $c_k(w)$ denotes the cost $c(w, \xi = \tilde{z}_k)$ for all scenarios $k = 1, \ldots, K$, as defined earlier in Section 3.2.

Before we investigate the Rademacher complexities, we first demonstrate the Lipschitz property of the regularized optimal solution mapping $w_\rho(\cdot)$ for any positive parameter ρ, based on the following assumption regarding the Lipschitz property of the cost function $c(w)$ and the strong convexity constant of the decision regularization function $\phi(\cdot)$.

Assumption 2. *The cost function $c(\cdot)$ and the decision regularization function $\phi(\cdot)$ satisfy the following conditions:*

A. $c(\cdot)$ is L_c-Lipschitz with respect to the decision $w \in S$, i.e., it holds that $\|c(w_1) - c(w_2)\|_2 \leq L_c\|w_1 - w_2\|$ for all $w_1, w_2 \in S$.

B. The decision regularization function $\phi(\cdot)$ is a 1-strongly convex function on the compact set S.

Note that we use the ℓ_2 norm as the norm on the space of outputs of the cost functions $c(\cdot)$, while the norm on the space of decisions w remains the generic norm $\|\cdot\|$. The reason for focusing on the ℓ_2 norm is that we can apply the elegant vector contraction inequality of [99] when analyzing the Rademacher complexity. It is also worth mentioning that the Lipschitz condition in Assumption 2.A implies that the cost functions $c_k(\cdot)$ are uniformly L_c-Lipschitz, i.e., $\|c(w_1) - c(w_2)\|_\infty \leq L_c\|w_1 - w_2\|$.

Proposition 3.3.1.1 (Lipschitz Properties of $w_\rho(\cdot)$ and $c(\cdot)$)**.** *Suppose Assumption 2 holds. Then, for any $\rho > 0$, the optimal solution mapping $w_\rho(\cdot)$ is $(\frac{L_c}{\rho})$-Lipschitz:*

$$\|w_\rho(p) - w_\rho(p')\| \leq \frac{L_c}{\rho}\|p - p'\|_2, \qquad \forall p, p' \in \Delta_K. \tag{3.8}$$

Furthermore, $c(w_\rho(\cdot))$ is $(\frac{L_c^2}{\rho})$-Lipschitz:

$$\|c(w_\rho(p)) - c(w_\rho(p'))\|_\infty \leq \|c(w_\rho(p)) - c(w_\rho(p'))\|_2 \leq \frac{L_c^2}{\rho}\|p - p'\|_2, \qquad \forall p, p' \in \Delta_K. \tag{3.9}$$

The proof of this Proposition follows standard arguments of Nesterov's smoothing technique ([106]), and a related result with a similar proof style appears in [59].

Proof. Proof of Proposition 3.3.1.1 Let $p, p' \in \Delta_K$ be fixed. We let $h_\rho(\cdot, p) : S \to \mathbb{R}$ be defined by $h_\rho(w, p) := \sum_{k=1}^K p_k c_k(w) + \rho\phi(w)$. Since $\phi(\cdot)$ is a 1-strongly convex function, then $h_\rho(\cdot, p)$ is ρ-strongly convex and it holds for all $w \in S$ and $g \in \partial_w h_\rho(w, p)$ that

$$h_\rho(w', p) - h_\rho(w, p) \;\geq\; g^T(w' - w) + \frac{\rho}{2}\|w' - w\|^2 \qquad \forall w' \in S. \tag{3.10}$$

Since $w_\rho(p) = \arg\min_{w \in S} h_\rho(w, p)$, the first-order optimality condition implies there exists a subgradient $g \in \partial h(w_\rho(p), p)$ such that $g^T(w' - w_\rho(p)) \geq 0$ for all $w' \in S$. Applying this condition in (3.10) with $w \leftarrow w_\rho(p)$ $w' \leftarrow w_\rho(p')$ yields

$$h_\rho(w_\rho(p'), p) - h_\rho(w_\rho(p), p) \;\geq\; \frac{\rho}{2}\|w_\rho(p') - w_\rho(p)\|^2.$$

Switching the role of p and p' yields

$$h_\rho(w_\rho(p), p') - h_\rho(w_\rho(p'), p') \;\geq\; \frac{\rho}{2}\|w_\rho(p) - w_\rho(p')\|^2.$$

Adding the above two inequalities together yields

$$\begin{aligned}
\rho\|w_\rho(p) - w_\rho(p')\|^2 \;&\leq\; h_\rho(w_\rho(p), p') - h_\rho(w_\rho(p'), p') + h_\rho(w_\rho(p'), p) - h_\rho(w_\rho(p), p) \\
&=\; [h_\rho(w_\rho(p), p') - h_\rho(w_\rho(p), p)] - [h_\rho(w_\rho(p'), p') - h_\rho(w_\rho(p'), p)] \\
&=\; \sum_{k=1}^K (p'_k - p_k) c_k(w_\rho(p)) - \sum_{k=1}^K (p'_k - p_k) c_k(w_\rho(p')) \\
&=\; \sum_{k=1}^K (p'_k - p_k)(c_k(w_\rho(p)) - c_k(w_\rho(p'))) \\
&\leq\; \|p - p'\|_2 \|c(w_\rho(p)) - c(w_\rho(p'))\|_2 \\
&\leq\; L_c\|p - p'\|_2 \|w_\rho(p) - w_\rho(p')\|,
\end{aligned}$$

where the last inequality uses Assumption (2.A). Dividing by $\|w_\rho(p) - w_\rho(p')\|$ leads to (3.8), and combining the resulting inequality again with (2.A) yields (3.9). □

To establish the generalization bound for the ICEO risk, we rely on both regular single-variate and multi-variate Rademacher complexity. In the ICEO setting, given a class of policies Π, where $\pi : \mathcal{X} \to S$ for all $\pi \in \Pi$, we can apply generalization bounds that directly use the Rademacher complexity of the function class $c \circ \Pi$. Given a sample $\{(x_i, \xi_i)\}_{i=1}^n$ the *empirical Rademacher complexity* $\hat{\mathfrak{R}}_n(c \circ \Pi)$ of the function class $c \circ \Pi$ is defined by

$$\hat{\mathfrak{R}}_n(c \circ \Pi) \;:=\; \mathbb{E}_\sigma\left[\frac{2}{n}\sup_{g \in c\circ\Pi} \sum_{i=1}^n \sigma_i g(x_i, \xi_i)\right] = \mathbb{E}_\sigma\left[\frac{2}{n}\sup_{\pi \in \Pi} \sum_{i=1}^n \sigma_i c(\pi(x_i), \xi_i)\right],$$

where σ_i are independent random variables drawn from the Rademacher distribution, i.e. $\Pr(\sigma_i = +1) = \Pr(\sigma_i = -1) = \frac{1}{2}$ for all $i = 1, 2, \ldots, n$. The *expected Rademacher complexity* $\mathfrak{R}_n(c \circ \Pi)$ is then defined as the expectation of $\hat{\mathfrak{R}}_n(c \circ \Pi)$ with respect to the i.i.d. sample $\{(x_i, \xi_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$:

$$\mathfrak{R}_n(c \circ \Pi) = \mathbb{E}_{(x_i,\xi_i)\sim\mathcal{D}}[\hat{\mathfrak{R}}_n(c \circ \Pi)].$$

Next, we introduce the multivariate Rademacher complexity as a generalization of the regular Rademacher complexity to a class of vector-valued functions. In the ICEO context, we focus on the hypothesis class $\mathcal{H}$ which takes values in Δ_K. Following [18], [99] and [9], the *empirical multivariate Rademacher complexity* $\hat{\mathfrak{R}}_n(\mathcal{H})$ is defined in our context as

$$\hat{\mathfrak{R}}_n(\mathcal{H}) = \mathbb{E}_\sigma \left[\frac{2}{n} \sup_{f\in\mathcal{H}} \sum_{i=1}^{n} \sum_{k=1}^{K} \sigma_{ik} f_k(x_i) \right],$$

where σ_{ik} are also independent random variables drawn from the Rademacher distribution for all $i = 1, 2, \ldots, n$ and $k = 1, \ldots, K$. Correspondingly, the *expected multivariate Rademacher complexity* $\mathfrak{R}_n(\mathcal{H})$ is then defined as

$$\mathfrak{R}_n(\mathcal{H}) = \mathbb{E}_{x_i\sim\mathcal{D}_x}[\hat{\mathfrak{R}}_n(\mathcal{H})],$$

In the remainder of this section, we provide generalization bounds with respect to the expected single-variate and multi-variate Rademacher complexities. We note that similar results can be achieved with respect to the empirical versions of the Rademacher complexities. Our focus on the expected versions is justified since, for many hypothesis classes $\mathcal{H}$, we can bound $\mathfrak{R}_n(\mathcal{H})$ by a term that converges to 0 as the sample size n grows. For example, [9] establish upper bounds of $\mathfrak{R}_n(\mathcal{H})$ for regularized linear hypothesis classes with the rate of $\mathcal{O}(\frac{1}{\sqrt{n}})$, where the $\mathcal{O}(\cdot)$ notation hides dimension dependent constants that depend on the type of regularization used.

Given a sample $\{(x_i, \xi_i)\}_{i=1}^n$, we aim to provide a high-probability bound on the out-of-sample risk $R(w_{\rho_n} \circ f)$, given the in-sample risks $\hat{R}_n(w_{\rho_n} \circ f)$ and $\hat{R}_n(w_{\rho_n} \circ f; \rho_n)$, that holds uniformly for any hypothesis $f \in \mathcal{H}$. As such, our generalization bound is constructed based on the the classic generalization bound with Rademacher complexity due to [11], which we restate below as specialized to the ICEO setting. Recall that $\bar{c} := \sup_{w\in S, \xi\in\Xi} c(w, \xi)$.

Theorem 3.3.2 ([11])**.** *Let Π be a family of functions mapping from $\mathcal{X}$ to S with bounded Rademacher complexity $\mathfrak{R}_n(c \circ \Pi)$. Then, for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over i.i.d. data $\{(x_i, \xi_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$, the following inequality holds for all $\pi \in \Pi$:*

$$R(\pi) \;\leq\; \hat{R}_n(\pi) + \mathfrak{R}_n(c \circ \Pi) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}}.$$

The next step of our analysis is to apply the vector contraction inequality of [99] to derive a generalization bound that depends directly on the multi-variate Rademacher complexity of the hypothesis class $\mathcal{H}$.

Theorem 3.3.3 (Generalization of ICEO)**.** *Suppose Assumption 2 holds and that the hypothesis class $\mathcal{H}$ has bounded multi-variate Rademacher complexity $\mathfrak{R}_n(\mathcal{H})$. Then, for any $\delta \in (0, 1]$ and $\rho_n > 0$, with probability at least $1 - \delta$ over i.i.d. data $\{(x_i, \xi_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$, the following inequalities hold for all $f \in \mathcal{H}$:*

$$\begin{aligned} R(w_{\rho_n} \circ f) &\leq \hat{R}_n(w_{\rho_n} \circ f) + \frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H}) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \\ &\leq \hat{R}_n(w_{\rho_n} \circ f; \rho_n) + \frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H}) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}}. \end{aligned} \tag{3.11}$$

Note that the right hand side of the first inequality in Theorem 3.3.3 involves the unregularized empirical risk, which may be evaluated for any $f \in \mathcal{H}$. The right hand side of the second inequality in Theorem 3.3.3 involves the regularized empirical risk, which is precisely the objective function of (ICEO-ρ). As mentioned previously, one can often establish upper bounds on $\mathfrak{R}_n(\mathcal{H})$ that converge to zero, for example at the rate $\mathcal{O}(\frac{1}{\sqrt{n}})$. Therefore, Theorem 3.3.3 suggests that we should set the sequence of regularization parameters ρ_n so that $\mathfrak{R}_n(\mathcal{H})/\rho_n$ converges to zero as well, in which case the remainder terms on the right-hand side of (3.11) converge to zero. We now state the proof of Theorem 3.3.3.

Proof. Proof of Theorem 3.3.3 Due to Proposition 3.3.1.1, in particular the Lipschitz property of $c(\cdot)$ in (3.9), we can apply the vector contraction inequality from [99] which, stated in terms of empirical Rademacher complexities, yields

$$\hat{\mathfrak{R}}_n(c \circ w_{\rho_n} \circ \mathcal{H}) \leq \frac{\sqrt{2}L_c^2}{\rho_n}\hat{\mathfrak{R}}_n(\mathcal{H}).$$

Taking expectations of both sides of the above inequality, with respect to i.i.d. data $\{(x_i, \xi_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$, yields

$$\mathfrak{R}_n(c \circ w_{\rho_n} \circ \mathcal{H}) \leq \frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H}).$$

Then, a direct application of Theorem 3.3.2 yields the first inequality. Finally, for any $\rho_n > 0$, note that $R(w_{\rho_n} \circ f) \leq R(w_{\rho_n} \circ f; \rho_n)$ due to the non-negativity of the decision regularization function $\phi(\cdot)$, which yields the second inequality. □

Leveraging Theorem 3, we obtain the following corollary that provides an upper bound for the optimality gap, in terms of out-of-sample risk evaluation, for the ICEO minimizer $\hat{f}^n_{\rho_n}$.

Corollary 3.3.3.1. *Suppose Assumption 2 holds and that the hypothesis class $\mathcal{H}$ has bounded multi-variate Rademacher complexity $\mathfrak{R}_n(\mathcal{H})$. Then, for any $\delta \in (0,1]$ and $\rho_n > 0$, with probability at least $1-\delta$ over i.i.d. data $\{(x_i,\xi_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$, the following inequality holds:*

$$R(w_{\rho_n} \circ \hat{f}^n_{\rho_n}) - R(w_{\rho_n} \circ f^*) \;\leq\; \frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H}) + \rho_n\bar{\phi} + \frac{3\bar{c}}{2}\sqrt{\frac{2\log(\frac{2}{\delta})}{n}},$$

where $\bar{\phi} := \sup_{w\in S}\phi(w)$ is the upper bound of the regularization term which is independent of the sample size.

Remark. *The generalization bound provided in Theorem 3.3.3 involves the Rademacher complexity $\mathfrak{R}_n(\mathcal{H})$ of the hypothesis class $\mathcal{H}$, which returns outputs in Δ_K. It is noteworthy that, if $\mathcal{H}$ involves a softmax operator, i.e., $\mathcal{H} = \text{soft} \circ \tilde{\mathcal{H}}$ for some hypothesis class $\tilde{\mathcal{H}}$ outputting in $\mathbb{R}^K$, then it is possible to obtain an identical generalization bound involving the Rademacher complexity $\mathfrak{R}_n(\tilde{\mathcal{H}})$ of the hypothesis class $\tilde{\mathcal{H}}$. Indeed, since the softmax function* soft *is Lipschitz ([55]) with constant 1, the Lipschitz bound (3.9) can be extended to a bound involving the outputs of $\tilde{\mathcal{H}}$ with the same (L_c^2/ρ) constant. Alternatively and equivalently, the vector contraction inequality of [99] can be extended to a Lipschitz mapping (such as the softmax function) that has both multivariate input and output. This extended vector contraction inequality can be obtained by a straightforward extension of Lemma 7 and Theorem 3 of [99] to the multivariate case.*

3.4 Computational Methods

In this section, we discuss the computational difficulties of solving the ICEO formulation (ICEO-ρ) and present multiple approaches to address them.

3.4.0.0.1 Non-convexity. First, we point out that the ICEO formulation, (ICEO-ρ), is not a convex optimization problem even in a very simple case where both the objective and constraints of the nominal optimization problem are linear and the decision regularization is quadratic, as stated in Example 3.4.1.

Example 3.4.1 (Linear nominal optimization problem)**.** Consider an example with a linear objective function in the optimization stage, i.e., $c_j(w)$ is a linear function $c_j^T w$ for some $c_j \in \mathbb{R}^d$ for all $j = 1,\ldots,K$. Suppose we use the decision regularization function $\phi(w) := \frac{1}{2}\|w\|_2^2$. For any $p \in \Delta_K$, let $\bar{c}(p) := \sum_{j=1}^K p_j c_j$. Then, note that

$$w_\rho(p) = \arg\min_{w\in S}\left\{\bar{c}(p)^T w + \tfrac{\rho}{2}\|w\|_2^2\right\} = \arg\min_{w\in S}\left\{\tfrac{\rho}{2}\|(\bar{c}(p)/\rho) - w\|_2^2\right\} = \Pi_S(\bar{c}(p)/\rho),$$

where $\Pi_S(\cdot)$ is the Euclidean projection operator onto S. Then, (ICEO-ρ) is the problem of minimizing a sum of linear functions composed with projection operators, which is generally non-convex. At best, when S is a polyhedron, i.e., $S := \{w \in \mathbb{R}^d : Aw \leq b\}$ and when we adopt a linear hypothesis class $\mathcal{H} = \{x \mapsto Bx \in \Delta_K : B \in \mathbb{R}^{K\times p}\}$, we can formulate (ICEO-ρ) as a bilinear quadratic optimization problem. Indeed, (ICEO-ρ) can be reformulated as

$$\begin{aligned}
\min_{B,w_i,\lambda_i} \quad & \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}(\mathbb{1}\{\xi_i = \tilde{z}_j\}(Bx_i)_j c_j^T w_i + (\rho/2)w_i^T w_i) && \text{(ICEO-}\rho\text{-LP)}\\
\text{s.t.} \quad & \frac{\rho}{2}w_i^T w_i + \sum_{j=1}^{K}(Bx_i)_j c_j^T w_i + \frac{1}{2}(\sum_{j=1}^{K}(Bx_i)_j c_j + A^T\lambda)^T(\sum_{j=1}^{K}(Bx_i)_j c_j + A^T\lambda))\\
& + \lambda_i^T b \leq 0, \quad \forall i = 1,\ldots,n\\
& Aw_i \leq b, \quad \forall i = 1,\ldots,n\\
& \lambda_i \geq 0, \quad \forall i = 1,\ldots,n
\end{aligned}$$

Note that the dual function of the nominal quadratic optimization is

$$-\frac{1}{2}(\sum_{j=1}^{K}(B^T x_i)_j c_j + A^T\lambda)^T(\sum_{j=1}^{K}(B^T x_i)_j c_j + A^T\lambda)) - \lambda^T b$$

thus the dual problem becomes

$$\min_{\lambda\geq 0} \frac{1}{2}(\sum_{j=1}^{K}(B^T x_i)_j c_j + A^T\lambda)^T(\sum_{j=1}^{K}(B^T x_i)_j c_j + A^T\lambda)) + \lambda^T b.$$

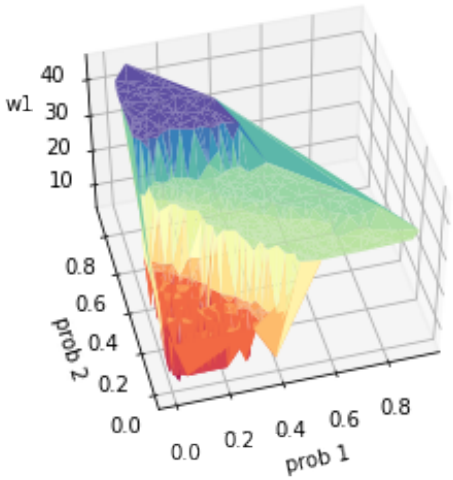
The first two group of constraints in (ICEO-ρ-LP) is to guarantee that w_i and λ_i are the optimal primal and dual solutions. The second and third group of constraints are for the primal and dual feasibility.

As demonstrated above, even in this simplest case, (ICEO-ρ-LP) is not a convex optimization problem. Besides non-convexity, a more serious issue from a practical standpoint is the potential of non-differentiability of the optimal solution mapping.

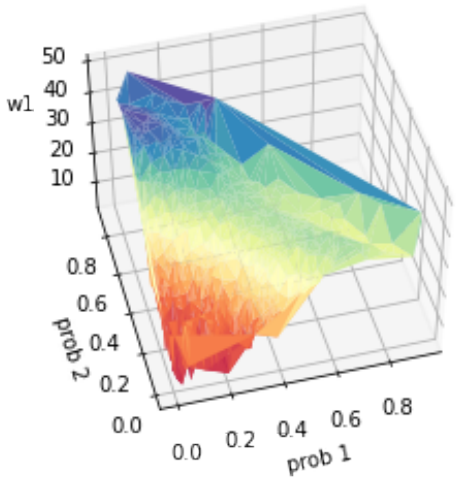
3.4.0.0.2 Non-differentiability. To solve the non-convex ICEO problem (ICEO-ρ), a default approach in machine learning is to use a gradient-based algorithm such as the basic stochastic gradient descent method. Indeed, in practice gradient-based algorithms are often able to deliver high quality solutions for machine learning problems, especially in high dimensions. Unfortunately, applying these basic gradient-based methods to solve the ICEO formulation poses an additional major difficulty due to the non-differentiability of the optimal solution mapping $w_\rho(\cdot)$. Although $w_\rho(\cdot)$ is a continuous function, as guaranteed for

example by Proposition 3.3.1.1, it is generally not differentiable. The non-differentiability leads to major difficulties in applying gradient-based method while solving ICEO-ρ. As reviewed in Section 3.1.1, existing studies that focused on directly learning the optimal solution mapping $w(f^*(x))$ also encounter the same issue of non-differentiability. [150] does not discuss much about this. [39] use the output of an automatic gradient function calculated by back propagation of a neural network. [2] approximate the gradient by solving a group of linear equations based on KKT conditions. However, all existing methods fail to demonstrate theoretical reliability or performance guarantees in approximating the gradient.

The non-differentiability of the optimal solution mapping mainly arises from the constraints and the points of discontinuity occur where there is a "jump" in the optimal solution, e.g., in the polyhedral case as in Example 3.4.1. Therefore, a non-differentiable optimal solution map may also have regions where it is constant (or close to constant), resulting in the gradient of the ICEO objective being equal to zero. We demonstrate this poor behavior in Figure 3.1a, where we plot the second coordinate of the optimal solution mapping $w_\rho(\cdot)$ with respect to the first two coordinates of the input probability vector, for the multi-product newsvendor problem in Example 3.2.1, demonstrating the piece-wise constant shape. Such piece-wise constant shapes will greatly impede the performance of gradient-based methods, even if the gradient is easily calculated. This is because the gradient of the optimal solution mapping is zero in flat regions creating poor local minima that are very difficult to escape.



(a) 3-D plot of the optimal oracle. It is clear that the landscape of the optimal solution mapping has cliffs and platforms.



(b) 3-D plot of the approximated oracle constructed using polynomial functions. The piece-wise constant shape is smoothed out.

Figure 3.1: The landscape of the optimal and the approximated oracle.

To address the issue of non-differentiability and its consequences leading to poor local optima and slow convergence, we develop a framework for approximating the mapping $w_\rho(\cdot)$

with a differentiable function $\tilde{w}_\rho(\cdot)$, which allows us to smooth out the optimal solution mapping and enhance convergence to a good local optimum. Figure 3.1b is an example of smoothing out the piece-wise constant shape by constructing an approximate oracle using polynomial kernel regression. As noted before, gradient-based methods are often highly effective at delivering high quality solutions to non-convex machine learning problems in practice. Thus, in a practical sense, the non-differentiability of the optimal solution mapping is a much more serious concern than the non-convexity. Our general strategy of approximating the optimal solution mapping with a differentiable function, for which we expand upon and give examples in Section 3.4.1, greatly increases the practical viability of the ICEO approach.

3.4.1 Approximate Optimal Solution Mappings

As stated in the previous section, the major computational difficulty in solving the ICEO training problem (ICEO-ρ) in practice is the non-differentiability of the mapping $w_\rho(\cdot)$. To overcome this difficulty, for any given ρ, we approximate the function $w_\rho(\cdot)$ with a differentiable function $\tilde{w}_\rho(\cdot) : \Delta_K \to S$. Then instead of (ICEO-$\rho$), we solve the following problem:

$$\begin{aligned} \min_{f \in \mathcal{H}} \quad & \frac{1}{n} \sum_{i=1}^{n} c(w_i, \xi_i) + \rho\phi(w_i) \\ \text{s.t.} \quad & w_i = \tilde{w}_\rho(f(x_i)) \end{aligned} \qquad \text{(Approx-ICEO-}\rho\text{)}$$

To construct such an approximation $\tilde{w}_\rho(\cdot)$, we rely on the ability to evaluate the optimal solution mapping $w_\rho(p)$ for any given $p \in \Delta_K$, as stated in Section 3.2. We can then generate a sequence of samples $(p_i, w_\rho(p_i))$ and build an approximation function $\tilde{w}_\rho(\cdot)$ using any class of continuous functions with enough representation power, such as polynomial functions or neural networks.

We consider two generic types of approximation schemes for building the mapping $\tilde{w}_\rho(\cdot)$: *(i)* uniform approximations, and *(ii)* high-probability approximations. Uniform approximation schemes satisfy a uniform error bound, as formalized below in Assumption 3, and can be achieved by an interpolation method such as the Bernstein polynomial method as described in Section 3.4.2.1. Note that, for each $j = 1, \ldots, K$ and $p \in \Delta_K$, we use the notation $w_{\rho,j}(p)$ and $\tilde{w}_{\rho,j}(p)$ to refer to the j^{th} coordinates of $w_\rho(p)$ and $\tilde{w}_\rho(p)$, respectively.

Assumption 3 (Uniform Error Bound)**.** *For each $j = 1, \ldots, K$, there exists a constant $\mathcal{E}_j^{\text{unif}} \geq 0$ such that the approximate optimal solution mapping $\tilde{w}_\rho(\cdot) : \Delta_K \to S$ satisfies:*

$$|w_{\rho,j}(p) - \tilde{w}_{\rho,j}(p)| \;\leq\; \mathcal{E}_j^{\text{unif}}, \qquad \forall p \in \Delta_K.$$

The uniform error bound in Assumption 3 provides guarantees for the approximation error over all probability vectors from the simplex Δ_K. There are two main drawbacks that apply to all known approaches for achieving a uniform error bound. First, achieving a

tight uniform error bound requires exact or near-exact computations of the optimal solution mapping $w_\rho(p)$ for all $p \in \Delta_K$. In practice, we may only have an approximate optimal solution mapping available. Second, the sample size required by a method that achieves Assumption 3, for example an interpolations scheme, may be prohibitively large. For these reasons we are motivated to consider a high-probability error bound, which would hold for the more realistic approach of using a regression method, possibly with noise in the output of $w_\rho(\cdot)$, to fit the approximate optimal solution mapping. We consider a generic approach that uses a hypothesis class $\mathcal{G}$ for the approximate optimal solution mappings. Assumption 4 below formalizes our high-probability error bound, which holds for a wide range of regression methods including, for example, the polynomial kernel regression method considered in Section 3.4.2.2. In Assumption 4, we work with a *reference distribution* $\mathcal{D}_p$ on Δ_K that we use to generate samples $\{p_i\}_{i=1}^m$ to feed into a regression method. In addition, for any $f \in \mathcal{H}$, we later use the notation $\mathcal{D}_{f(x)}$ to refer to the distribution on Δ_K induced by the marginal distribution $\mathcal{D}_x$ of $x \in \mathcal{X}$.

Assumption 4 (High-probability Error Bound)**.** *Let $\mathcal{G}$ be a family of candidate approximate optimal solution mappings whereby $\tilde{w}_\rho(\cdot) : \Delta_K \to S$ for all $\tilde{w}_\rho(\cdot) \in \mathcal{G}$. For each $j = 1, \ldots, K$, there exists a function $\mathcal{E}_j^{\text{prob}}(\cdot, \cdot; \mathcal{G}) : \mathbb{N} \times [0, 1) \to [0, \infty)$ such that, for any distribution $\mathcal{D}_p$ on Δ_K and for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ over m independent samples drawn from $\mathcal{D}_p$ with empirical distribution $\hat{\mathcal{D}}_p^m$, it holds for all $\tilde{w}_\rho(\cdot) \in \mathcal{G}$ that:*

$$\left| \mathbb{E}_{\mathcal{D}_p}[|w_{\rho,j}(p) - \tilde{w}_{\rho,j}(p)|] - \mathbb{E}_{\hat{\mathcal{D}}_p^m}[|w_{\rho,j}(p) - \tilde{w}_{\rho,j}(p)|] \right| \leq \mathcal{E}_j^{\text{prob}}(m, \delta; \mathcal{G}).$$

When using an approximate optimal solution mapping with either a uniform or a high-probability error bound guarantee, a natural question is: do the performance guarantees of the ICEO approach developed in Section 3.3 extend to problem (Approx-ICEO-ρ)? We now answer this question affirmatively by extending the generalization bounds of Theorem 3.3.3 to situations with approximate mappings satisfying either Assumption 3 or Assumption 4. We make an implicit assumption that, after solving problem (Approx-ICEO-ρ), the decision-maker uses the correct optimal solution mapping $w_\rho(\cdot)$ to make decisions. Therefore, the left hand side of our bounds involve the true risk $R(w_\rho \circ f)$ with the correct mapping while the right hand sides involve the empirical risk $\hat{R}_n(\tilde{w}_\rho \circ f)$ with the approximation (and the reguarlized version thereof).

Theorem 3.4.1. *Suppose Assumption 2 holds and that the hypothesis class $\mathcal{H}$ has bounded multi-variate Rademacher complexity $\mathfrak{R}_n(\mathcal{H})$. Then, for any $\delta \in (0, 1]$ and $\rho_n > 0$, we have the following:*

(i) If the approximate optimal solution mapping $\tilde{w}_\rho(\cdot)$ satisfies the uniform error bound as stated in Assumption 3, then with probability at least $1 - \delta$ over i.i.d. data $\{(x_i, \xi_i)\}_{i=1}^n$

drawn from the distribution $\mathcal{D}$, the following inequalities hold for all $f \in \mathcal{H}$:

$$R(w_{\rho_n} \circ f) \leq \hat{R}_n(\tilde{w}_{\rho_n} \circ f) + \frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H}) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} + L_c\sum_{j=1}^{d}\mathcal{E}_j^{\text{unif}} \tag{3.12}$$

$$\leq \hat{R}_n(\tilde{w}_{\rho_n} \circ f; \rho_n) + \frac{\sqrt{2}L_c^2}{\rho_n}\mathfrak{R}_n(\mathcal{H}) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} + L_c\sum_{j=1}^{d}\mathcal{E}_j^{\text{unif}}. \tag{3.13}$$

(ii) If the approximate optimal solution mapping $\tilde{w}_\rho(\cdot)$ comes from a family $\mathcal{G}$ satisfying the high probability error bound as stated in Assumption 4, then with probability at least $1-\delta$ over i.i.d. data $\{(x_i, \xi_i)\}_{i=1}^n$ drawn from the distribution $\mathcal{D}$ and over m independent samples $\{p_i\}_{i=1}^m$ drawn from a reference distribution $\mathcal{D}_p$ on Δ_K, the following inequalities hold for all $f \in \mathcal{H}$:

$$R(w_{\rho_n} \circ f) \leq \hat{R}_n(\tilde{w}_{\rho_n} \circ f) + L_c\sum_{j=1}^{d}\left[\frac{1}{m}\sum_{i=1}^{m}|w_{\rho,j}(p_i) - \tilde{w}_{\rho,j}(p_i)| + \mathcal{E}_j^{\text{prob}}(n, \delta/2d; \mathcal{G}) + \mathcal{E}_j^{\text{prob}}(m, \delta/2d; \mathcal{G})\right]$$
$$+ \frac{\sqrt{2}L_c^2}{\rho}\mathfrak{R}_n(\mathcal{H}) + \text{diam}(S)L_c\text{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \tag{3.14}$$

$$\leq \hat{R}_n(\tilde{w}_{\rho_n} \circ f; \rho_n) + L_c\sum_{j=1}^{d}\left[\frac{1}{m}\sum_{i=1}^{m}|w_{\rho,j}(p_i) - \tilde{w}_{\rho,j}(p_i)| + \mathcal{E}_j^{\text{prob}}(n, \delta/2d; \mathcal{G}) + \mathcal{E}_j^{\text{prob}}(m, \delta/2d; \mathcal{G})\right]$$
$$+ \frac{\sqrt{2}L_c^2}{\rho}\mathfrak{R}_n(\mathcal{H}) + \text{diam}(S)L_c\text{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}}. \tag{3.15}$$

3.4.2 Approximate the Optimal Solution Oracle by Polynomials

In this section, we provide two examples of using polynomial functions to approximate the optimal solution mapping: *(i)* interpolation using Bernstein polynominals, which satisfies a uniform error bound, and *(ii)* polynomial kernel regression, which satisfies a high-probability error bound.

3.4.2.1 Bernstein polynomials.

One example for approximating the optimal solution mapping is interpolation using Bernstein polynomials, for which we review the definition below.

Definition 3.4.1 (**Bernstein approximation ([38])).** For a given function $\omega : \Delta_K \to \mathbb{R}$, the Bernstein approximation with order s, $B_s(\omega) : \Delta_K \to \mathbb{R}$, is defined by:

$$B_s(\omega)(p) := \sum_{\alpha \in I(K,s)} \bar{w}\left(\frac{\alpha}{s}\right)\frac{s!}{\alpha!}p^\alpha, \quad \forall p \in \Delta_K,$$

where $I(K,s) := \{\alpha \in \mathbb{N}_0^K \,|\, \sum_{i=1}^K \alpha_i = s\}$, $\alpha! := \Pi_i \alpha_i!$, and $p^\alpha := p_1^{\alpha_1} \cdots p_K^{\alpha_K}$.

Using Bernstein polynomials, based on a result of [38], we can achieve a uniform bound of the approximation error as described in Assumption 3.

Proposition 3.4.1.1. *For a given $\rho > 0$, suppose that we use the Bernstein approximation method (Definition 3.4.1) applied separately to each coordinate function $w_{\rho,j}(\cdot)$ to construct an approximate optimal solution mapping $\tilde{w}_\rho(\cdot)$. Then, $\tilde{w}_\rho(\cdot)$ satisfies the uniform error bound in Assumption 3 with $\mathcal{E}_j^{\text{unif}} = \frac{\Omega L_c}{\rho\sqrt{s}}$, where $\Omega > 0$ is an absolute constant.*

Proof. Proof of Proposition 3.4.1.1 This result directly follows from Theorem 3.2 in [38] together with the Lipschitz property from Proposition 3.3.1.1. □

Given the result in Proposition 3.4.1.1, we can immediately obtain a generalization bound for the Bernstein approximation method by applying item *(i)* of Theorem 3.4.1. While the Bernstein polynomial method provides a strong uniform error bound guarantee, there is a significant drawback in the number of samples required to obtain this bound. Indeed, to accomplish this approximation, it involves knowing function values of $w_{\rho,j}(\cdot)$ on the grid $\Delta_{K,s} := \{w \in \Delta_K : sw \in \mathbb{N}_0^K\}$ which has $\binom{K+s}{K}$ many points in total. As such, the number of calculations of $w_\rho(\cdot)$ may be prohibitively large, which motivates the use of regression methods.

3.4.2.2 Polynomial kernel regression.

In this section, we consider using the less computationally prohibitive regression methods that lead to high-probability bounds as in Assumption 4. As an exemplary case, we consider the polynomial kernel regression method. In this setting, we allow for the possibility of a "noised oracle" whereby the optimal solution mapping is not computed exactly. Specifically, the noised oracle outputs $w_\rho(p) + \sigma\varepsilon$ instead of $w_\rho(p)$, where ε is a d-dimensional standard Gaussian random vector and σ is a scalar that represents the standard deviation of the noise. The approximate oracle $\tilde{w}_\rho(\cdot)$ is constructed on independent samples $\{(p_i, w_i)\}_{i=1}^m$ where p_i is drawn from the reference distribution $\mathcal{D}_p$ and w_i is computed from the noised oracle. That is, we assume that $w_i = w_\rho(p_i) + \sigma\varepsilon_i$ for $\{\varepsilon_i\}_{i=1}^m$ that are i.i.d. realizations of Gaussian random variables. These samples can be achieved by first generating $\{p_i\}_{i=1}^m$ randomly from following any user-chosen distribution $\mathcal{D}_p$ over the simplex Δ_K and then calculating $\{w_i\}_{i=1}^m$ from a (possibly randomized) algorithm for approximating $w_\rho(\cdot)$. Note that we do assume that the noise is Gaussian, which may be a reasonable assumption for some algorithmic schemes for approximating $w_\rho(\cdot)$.

The approximate optimal solution mapping is learned using polynomial kernels $k(p, p') = (c + p^T p')^s$, where $s \in \mathbb{N}$ is the degree parameter. In the remaining part of this section, we let $\mathcal{G}$ denote a function class induced by a polynomial kernel of degree s and let $\|\cdot\|_\mathcal{G}$ denote any norm defined on $\mathcal{G}$. Note that $\mathcal{G}$ is a convex, star-shaped function class [143]. For the function class $\mathcal{G}$ and a given sample $\{p_i\}_{i=1}^m$, let $\tau_j(\mathcal{G}, \{p_i\}_{i=1}^m, r) := \inf_{u \in \mathcal{G}: \|u\|_\mathcal{G} \le r}(\frac{1}{m}\sum_{i=1}^m (u(p_i) -$

$w_{\rho,j}(p_i))^2)^{1/2}$ denote the fitting ability for $w_{\rho,j}$ using the kernel function class $\mathcal{G}$ within a user-defined radius r. Given the function class $\mathcal{G}$ and a given sample $\{(p_i, w_i)\}_{i=1}^m$, the method of kernel ridge regression estimates the approximate optimal solution mapping $\tilde{w}_\rho(\cdot)$ by solving:

$$\min_{u\in\mathcal{G}:\|u\|_{\mathcal{G}}\leq r} \frac{1}{m}\sum_{i=1}^{m}(u(p_i) - w_{i,j})^2 \tag{3.16}$$

The corresponding high-probability approximation error bound for learning the noised oracle using polynomial kernel ridge regression.

Proposition 3.4.1.2. *Let $\mathcal{G}$ denote a function class induced by a polynomial kernel of degree s, suppose that the noise of the output has standard deviation σ, and that we construct the approximate solution mapping $\tilde{w}_\rho(\cdot)$ using kernel ridge regression (3.16) with a user-defined radius $r > 0$. Then, there exist absolute constants $\bar{c}, \bar{c}'$ such that for all $\delta_m \geq \bar{c}_0 \frac{\sigma}{r}\frac{(s-1+K)!}{(s-1)!K!}\frac{1}{m}$, we have*

$$\frac{1}{m}\sum_{i=1}^{m}(\tilde{w}_{\rho,j}(p_i) - w_{\rho,j}(p_i))^2 \;\leq\; \bar{c}_1(\bar{c}_1'\tau_j(\mathcal{G}, \{p_i\}_{i=1}^m, r) + r^2\delta_m^2),$$

with probability at least $1 - \bar{c}_2\exp(-\bar{c}_2'\frac{mr^2}{\sigma^2}\delta_m^2)$ for each coordinate $j = 1, \ldots, K$. Moreover, for any θ_m that satisfies $\theta_m \geq \bar{c}_3\sqrt{\frac{1}{m}\frac{(s-1+K)!}{(s-1)!K!}}$, if it also holds that $m\theta_m^2 \geq \bar{c}_0\log(4\log(\frac{1}{\theta_m}))$, then

$$\left|\mathbb{E}_p[(\tilde{w}_{\rho,j}(p) - w_{\rho,j}(p))^2]^{1/2} - \left(\frac{1}{m}\sum_{i=1}^{m}(\tilde{w}_{\rho,j}(p_i) - w_{\rho,j}(p_i))^2\right)^{1/2}\right| \;\leq\; \bar{c}_3 r^2\theta_m$$

with probability at least $1 - \bar{c}_4\exp(-\bar{c}_4'\frac{m\theta_m^2}{r^2})$ for each coordinate $j = 1, \ldots, K$.

The main body of the proof is a generalization of the result in Example 13.19 of [143]. We have the corresponding generalization bound in the following corollary.

Corollary 3.4.1.1. *Suppose Assumption 2 holds and that the hypothesis class $\mathcal{H}$ has bounded multi-variate Rademacher complexity $\mathfrak{R}_n(\mathcal{H})$. Suppose further that we employ kernel ridge regression (3.16) using a function class $\mathcal{G}$ induced by a polynomial kernel of degree s under the same conditions as in Proposition 3.4.1.2. Then, for any $\delta \in (0, 1]$ and $\rho_n > 0$, the*

following inequalities hold for all $f \in \mathcal{H}$*:*

$$\begin{aligned} R(w_{\rho_n} \circ f) \;\leq\; & \hat{R}_n(\tilde{w}_{\rho_n} \circ f) + L_c \sum_{j=1}^{d} [\bar{c}_3 r^2(\theta_m + \theta_n) + \tau_j(\mathcal{G}, \{p_i\}_{i=1}^m, r) + \bar{c}_1' r \delta_m] \\ & + \frac{\sqrt{2} L_c^2}{\rho_n} \mathfrak{R}_n(\mathcal{H}) + L_c \bar{w}_j \sqrt{2\mathrm{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p)} + \bar{c} \sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \\ \leq\; & \hat{R}_n(\tilde{w}_{\rho_n} \circ f; \rho_n) + L_c \sum_{j=1}^{d} [\bar{c}_3 r^2(\theta_m + \theta_n) + \tau_j(\mathcal{G}, \{p_i\}_{i=1}^m, r) + \bar{c}_1' r \delta_m] \\ & + \frac{\sqrt{2} L_c^2}{\rho_n} \mathfrak{R}_n(\mathcal{H}) + L_c \bar{w}_j \sqrt{2\mathrm{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p)} + \bar{c} \sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \end{aligned}$$

with probability at least $1 - \delta'$ *over i.i.d. data* $\{(x_i, \xi_i)\}_{i=1}^n$ *drawn from the distribution* $\mathcal{D}$ *and over* m *independent samples* $\{(p_i, w_i)\}_{i=1}^m$*, where* $\delta' = \delta + \bar{c}_2 \exp(-\bar{c}_2' \frac{mr^2}{\sigma^2} \delta_m^2) + \bar{c}_4(\exp(-\bar{c}_4' \frac{m\theta_m^2}{r^2}) + \exp(-\bar{c}_4' \frac{n\theta_n^2}{r^2}))$ *and* δ_m, θ_m*, and* θ_n *are chosen to satisfy the conditions in Proposition 3.4.1.2.*

3.4.3 Computational Methods for the Semi-algebraic Case

In this section, we present an approach based on polynomial optimization in the case where the objective of the downstream optimization problem is semi-algebraic and we use a linear hypothesis class. In this case, when we additionally use a polynomial approximation $\tilde{w}_\rho(\cdot)$, we can reformulate the approximate ICEO formulation (Approx-ICEO-ρ) as a polynomial optimization problem, which can be solved with a hierarchy of semi-definite optimizaiton problems. Specifically we assume that both c and ϕ are semi-algebraic functions and we consider the linear hypothesis class $\mathcal{H} = \{f(x) : f(x) = Bx + b, (B, b) \in \mathcal{B}\}$ where $\mathcal{B}(\mathcal{X}) = \{(B, b) \in \mathbb{R}^{K \times p} \times \mathbb{R}^K : f(x) \in \Delta_K \; \forall x \in \mathcal{X}\}$ ensures that the output of the hypothesis returns a feasible probability vector. In this section, we demonstrate an exact solution method for the semi-algebraic case by transforming the (Approx-ICEO-ρ) to a polynomial optimization program. Before we reach the reformulated problem, we first review the definitions of semi-algebraic sets and semi-algebraic functions.

Definition 3.4.2 (**Semi-algebraic set ([88])).** $K \subset \mathbb{R}^n$ is a basic semi-algebraic set if

$$K = \{x \in \mathbb{R}^n : g_j(x) \geq 0, j = 1, \ldots, m\}$$

for some polynomial functions $(g_j)_{j=1}^m$, i.e., $(g_j)_{j=1}^m \subset \mathbb{R}[x]$, where $\mathbb{R}[x]$ denotes the ring of real polynomials.

Similarly, a semi-algebraic set is defined by not only a finite sequence of polynomial inequalities, but also equations, or finite union of these.

Definition 3.4.3 (**Semi-algebraic function ([88]).** Let K be a semi-algebraic set of $\mathbb{R}^n$. A function $f : K \to \mathbb{R}$ is a semi-algebraic function if its graph $\Psi_f := \{(x, f(x)) : x \in K\}$ is a semi-algebraic set of $\mathbb{R}^n \times \mathbb{R}$.

Functions generated by finitely many of dyadic operations $\{+, \times, \div, \vee, \wedge\}$ and monadic operations $|\cdot|$ and $(\cdot)^{1/q}$, $q \in \mathbb{N}$, on polynomials are semi-algebraic.

Note that with the linear hypothesis class, we need an additional constraint $Bx+b \in \Delta_K$, to guarantee that the output $f(x)$ is a valid probability vector for any $x \in \mathcal{X}$. We also assume that $\mathcal{X}$ is a polyhedron, i.e. $\mathcal{X} := \{x \in \mathbb{R}^p : Ax \geq a\}$ for some $A \in \mathbb{R}^{m\times p}$ and $a \in \mathbb{R}^m$. Then the problem Approx-ICEO-ρ becomes:

$$\begin{aligned} \min_{B,b} \quad & \frac{1}{n}\sum_{i=1}^{n} c(w_i, \xi_i) + \frac{\rho}{2}\phi(w_i) && \text{(Poly-Approx-ICEO-}\rho_n\text{)} \\ \text{s.t.} \quad & w_i = \tilde{w}_\rho(Bx_i + b) \\ & Bx + b \in \Delta_K, \forall x \in \mathcal{X} = \{x \in \mathbb{R}^p : Ax \geq a\} \end{aligned}$$

Note that the approximated oracle $\tilde{w}_\rho(\cdot)$ is constructed by polynomial kernels, so the first group of constraints are polynomial functions. Then we show that the second group of constraints can be reformulated to a group of linear constraints using the following proposition.

Proposition 3.4.1.3. *Suppose $\mathcal{X} := \{x \in \mathbb{R}^p : Ax \geq a\}$ for some $A \in \mathbb{R}^{m\times p}$ and $a \in \mathbb{R}^m$, then the constraint*

$$Bx + b \in \Delta_K, \forall x \in \mathcal{X}$$

can be rewritten as the following group of constraints by introducing new decision variables $y_k \in \mathbb{R}^m, k = 1, \ldots, K$, $z, u \in \mathbb{R}^m$

$$\begin{cases} a^T y_k \geq -b_k \quad \forall k = 1, \ldots, K \\ A^T y_k = B_k \quad \forall k = 1, \ldots, K \\ a^T z \geq 1 - \mathbb{1}^T b \\ A^T z = B^T \mathbb{1} \\ a^T u \geq -1 + \mathbb{1}^T b \\ A^T u = -B^T \mathbb{1} \\ y_k, z, u \geq 0 \quad \forall k = 1, \ldots, K \end{cases}$$

We have now shown that problem (Poly-Approx-ICEO-ρ_n) is a problem optimizing a basic semi-algebraic function on a basic semi-algebraic set which, by Proposition 11.10 of [88], can be reformulated as a polynomial optimization problem, which can be solved by solving a hierarchy of semi-definite problems.

3.5 Numerical Experiments

In this section, we demonstrate the numerical performance of our proposed ICEO framework using synthetic data. We first summarize the benchmark methods that we adopted for comparison:

1. Sample average approximation (SAA). In this benchmark, the decision-maker simply ignores the contextual features and minimizes the average of cost functions using empirical distribution of the observations of the random parameter.

2. The two-step estimate-then-optimize (ETO) method based on cross-entropy loss (ETO-Entropy). In this benchmark, we estimate the hypothesis $f \in \mathcal{H}$ using the cross-entropy loss function (a standard loss function for multi-class classification) instead of the downstream optimization goal.

3. The prescriptive method (PRES) proposed by [18]. We consider the KNN-based (PRES-KNN) and kernel-based (PRES-Kernel) variants.

4. The stochastic optimization forest (SO-Forest) method proposed by [70].

3.5.0.0.1 Data Generation Process. The synthetic data is generated in the following manner. The features $x_i \in \mathbb{R}^p$ are generated independently following the multi-variate Gaussian distribution $x_i \sim N(0, MI_p)$ for some constant $M > 0$ and where I_p is an identity matrix. Then we consider $K = 4$ scenarios for the random parameter $\xi_i \in \mathbb{R}^2$, i.e., $\Xi = \{\tilde{z}_1, \tilde{z}_2, \tilde{z}_3, \tilde{z}_4\}$. Then, the corresponding conditional probability vector of ξ_i is generated according to $\text{soft}((B^* x_i + b^*)^{\text{deg}})$, with $B^* \in \mathbb{R}^{K \times p}$, $b^* \in \mathbb{R}^K$ and deg a positive integer being parameters set before the data generation process. Then, ξ_i takes the value of $\tilde{z}_k$ with probability $p_k^*(x) = \text{soft}((B^* x_i + b^*)^{\text{deg}})_k$ for all $k = 1, \ldots, K$.

3.5.0.0.2 Optimal Solution Mapping Approximation. The optimal oracle is approximated using neural networks in the experiment. We first generate a data set $\{(p_i, w_i)\}_{i=1}^m$ by uniformly sampling p_i from the simplex Δ_K and then generating $w_i := w_\rho(p_i)$. Then we train a neural network with one hidden layer to approximate the oracle. The neural network is trained with respect to the mean absolute percentage error (MAPE) loss.

3.5.0.0.3 ICEO Hypothesis Learning. In this experiment, we consider two candidate hypothesis classes for $\mathcal{H}$. First, we consider a softmax function composed with a linear function class, whereby $\mathcal{H}_1 := \text{soft} \circ \tilde{\mathcal{H}}_1$ and $\tilde{\mathcal{H}}_1 := \{x \mapsto Bx + b_0 : B \in \mathbb{R}^{K \times p}, b_0 \in \mathbb{R}^K\}$. The other case is $\mathcal{H}_2 := \text{soft} \circ \tilde{\mathcal{H}}_2$ where $\tilde{\mathcal{H}}_2$ denotes a neural network with one-hidden layer. When the degree parameter deg of the data generation process is higher than one, then there is a model misspecification when learning the ICEO hypothesis with $\mathcal{H}_1$ since the true hypothesis $f^*(x) := \text{soft}((Bx_i + b)^{\text{deg}})$ is not in the hypothesis class $\mathcal{H}_1$ when deg > 1. In

both cases, we solve (Approx-ICEO-ρ) using the Adam optimization algorithm ([78]) to learn the hypothesis.

3.5.1 Multi-item Newsvendor Problem

We consider the multi-item newsvendor problem, as in Example 3.2.1, with synthetic data. We consider $d = 2$, which is the case where the newsvendor jointly decides the order quantities of two products with an overall budget of 50. The decision variable $w \in \mathbb{R}^2$ and random demand $\xi \in \mathbb{R}^2$ are both two-dimensional, corresponding to the order quantity and demand of the two products. The newsvendor aims to minimize the total inventory cost as formulated in (3.4), with unit overstock costs h_1 and h_2 set to 1 and 1.3 and unit stockout cost b_1 and b_2 set to 9 and 8 for the two products, respectively.

3.5.1.0.1 Results and Comparisons with Benchmarks. In this experiment, we consider $\{\tilde{z}_1 := (33, 15), \tilde{z}_2 := (71, 4), \tilde{z}_3 := (17, 47), \tilde{z}_4 := (4, 43)\}$. In the data generation process, we set $M = 5$, deg $= 1$, and each element of B as an integer between 0 and 150. We set the regularization coefficient $\rho = 0.01$. We consider multiple training set sizes $n \in \{100, 300, 500, 700\}$ and for every value of n, we run 25 simulations. We use a validation set to tune the hyper-parameters for PRES-KNN, PRES-Kernel, ETO-Entropy and the ICEO method. For the SO-Forest method, we adopt the default setting provided in [70]. To evaluate out-of-sample performances of all these methods, we generate a test set including 1000 samples in each simulation. We would like to emphasize that we evaluate performance using unregularized ICEO objective (i.e., the newsvendor cost (3.4)) rather than the ICEO objective with regularization.

Figure 3.2 demonstrates the performance of the ICEO method and the non-parametric benchmarks. As demonstrated in this plot, the ICEO method outperforms all benchmarks with all training set sizes. The advantage of the ICEO method compared to the best benchmark (PRES-KNN) is more significant when the sample size is small ($n = 100$), which agrees with our intuition that nonparametric methods should perform better with more data. When compared with non-parametric prescriptive methods, our method shows the benefit of modeling the underlying conditional distribution.

3.5.1.0.2 Results on Model Misspecification We also investigate the case of model misspecification under the softmax linear hypothesis class $\mathcal{H}_1$. Since the ICEO method and ETO-Entropy are the only two methods that involves modeling the underlying conditional distribution using this hypothesis class, we compare the performance of just these two methods. To evaluate the performance of both methods, we consider the newsvendor cost on a test set with size 1000 in each simulation. To better quantify the improvement of ICEO compared to ETO-Entropy, we define the *relative improvement* as $\frac{\text{cost(ETO}-\text{Entropy)}-\text{cost(ICEO)}}{\text{cost(ETO}-\text{Entropy)}}$. In Figure 3.3a, we show the performance of the ICEO method compared to the two-step ETO-Entropy method and Figure 3.3b demonstrates the relative improvement of the ICEO

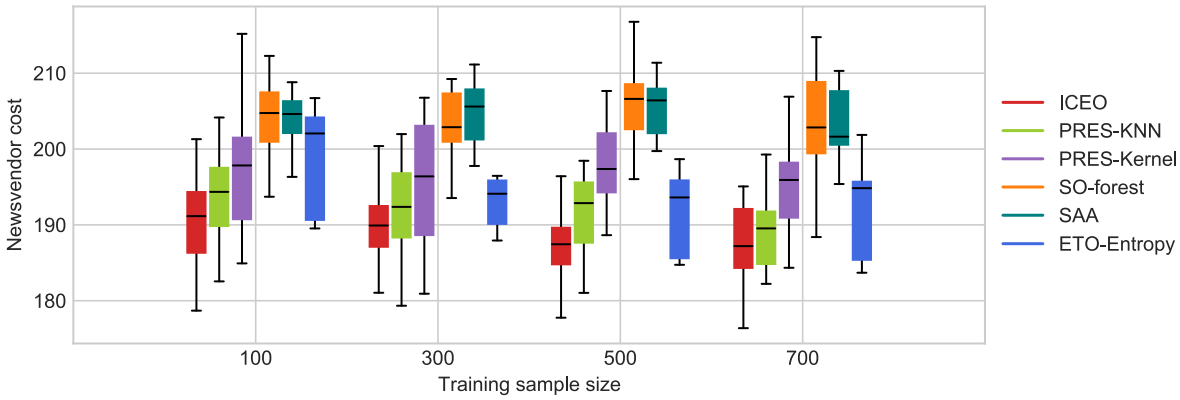


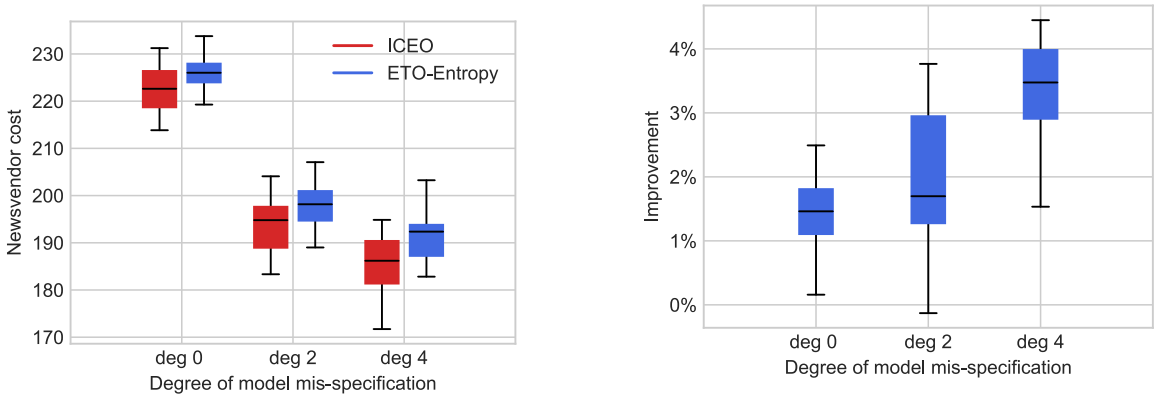
Figure 3.2: Comparison of ICEO with benchmark methods on the multi-item newsvendor problem.

method as compared to the ETO-Entropy method. As we can see, under model misspecification, ICEO consistently outperforms two-step ETO-Entropy method and the advantage increases when the degree of model misspecification increases. Both plots demonstrate the advantage of considering the ultimate optimization goal while estimating the conditional distribution. In addition, we note a decreasing trend of the out-of-sample cost for both methods in Figure 3.3a. Indeed, as the degree of model misspecification increase, the components in the probability vector become more concentrated on a single outcome. In other words, with a higher degree of model misspecification, the demand becomes more deterministic, which makes it easier for both methods to learn the conditional demand distribution and results in smaller cost values.

3.5.2 Quadratic Cost Network Flow Problem

In this part, we consider the minimum cost network flow problem. The problem formulation is as stated in Example 3.2.3, where $g_d(w_d, \xi_d) = c_d(w_d - \xi_d)^2$. We consider a simple network demonstrated in Figure 3.4, where there are two source nodes 1 and 2, and two sink nodes, 3 and 4. The amount of flow that sources out of each of the source nodes 1 and 2 must be no less than a threshold equal to 10. Similarly, the amount of flow that goes into each of the sink nodes 3 and 4 must be at least 10. We let w_1, w_2, w_3, w_4 denote the amount of flow on arcs $(1,3), (1,4), (2,3), (2,4)$, respectively.

3.5.2.0.1 Results and Comparisons with Benchmarks. In this experiment, we consider randomly generated $\{\tilde{z}_1 = (7.004, 8.442, 6.765, 7.279), \tilde{z}_2 = (9.515, 0.127, 4.136, 0.488), \tilde{z}_3 = (0.999, 5.081, 2.002, 7.442), \tilde{z}_4 = (1.929, 7.008, 2.932, 7.745)\}$. Furthermore, the weights of each arc c_1, c_2, c_3, c_4 take the values of $1, 3, 2, 2$ respectively. The regularization coefficient $\rho = 0.01$ and deg $= 1$. We consider ETO-Entropy, SAA, PRES-KNN, PRES-Kernel as



(a) Out-of-sample performance

(b) Improvement of ICEO compared to ETO-Entropy

Figure 3.3: Comparison between ICEO and ETO-Entropy under model misspecification on the multi-item newsvendor problem.

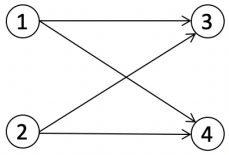


Figure 3.4: Network graph

benchmarks. We do not include SO-forest as a benchmark because there is no existing implementation of SO-forest for the quadratic cost network flow problem. As in Section 3.5.1, we consider multiple training set sizes $n \in \{100, 300, 500, 700\}$, we run 25 simulations for each sample size, and the test set includes 1000 samples in each simulation. To tune the hyper-parameters for ICEO, ETO-Entropy, and the two prescriptive methods, PRES-KNN and PRES-Kernel, we use a validation set including 1000 samples.

Figure 3.5 compares the test-set performance of the ICEO method and benchmarks. As demonstrated in this plot, the performance of ICEO method outperforms all benchmarks. Compared to the best benchmark (PRES-KNN), the advantage of ICEO is more significant when sample size is limited.

3.5.2.0.2 Results on Model Misspecification. As in Section 3.5.1, we investigate the case of model misspecification when considering the softmax linear hypothesis class $\mathcal{H}_1$. Similar to Section 3.5.1, we compare the test-set performances of ICEO and ETO-Entropy. Figure

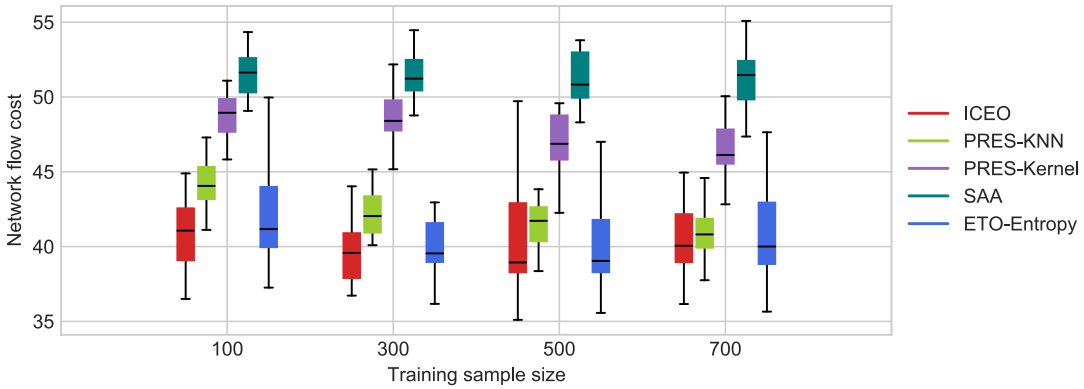
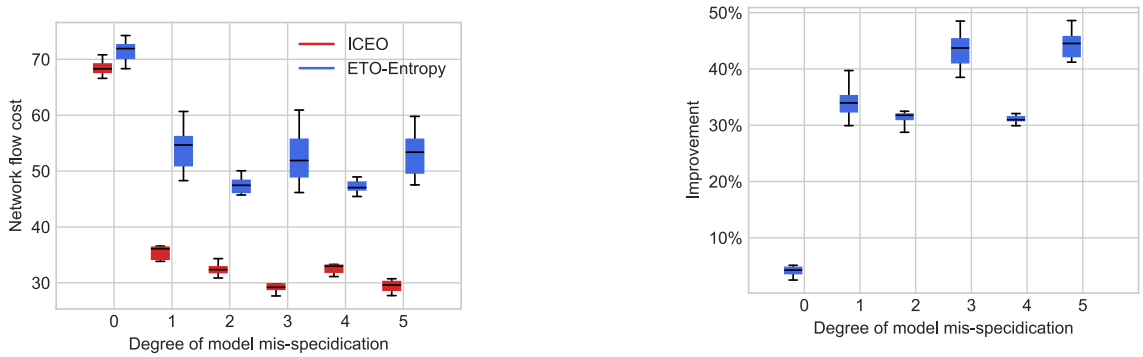


Figure 3.5: Comparison of ICEO with benchmark methods on the min-cost network flow problem.



(a) Out-of-sample performance

(b) Improvement of ICEO compared to ETO-Entropy

Figure 3.6: Comparison between ICEO and ETO-Entropy under model mis-specification on the min-cost network flow problem.

3.6a demonstrates the test-set performance and Figure 3.6b illustrates the improvement of ICEO method compared to the two-step benchmark, ETO-Entropy. In each configuration, we conducted 15 simulations. Figure 3.6a demonstrates that the ICEO method always outperforms the two-step benchmark Entropy. According to Figure 3.6b, there is a significant increase in the improvement of when model misspecification is introduced, which, again, demonstrates the advantage of integrating the optimization objective into the estimation stage. Figure 3.6b also demonstrates an appearing trend that the improvement of ICEO compared to Entropy is increasing as the degree of model misspecification increases.

3.6 Conclusion

In this paper, we propose a new framework for estimating the underlying conditional distribution in contextual stochastic optimization. The proposed ICEO framework uses a flexible hypothesis class to learn the hypothesis by incorporating the downstream optimization goal.

The ICEO framework developed herein applies to the case where the random parameter is a discrete random variable and the nominal optimization problem is convex. To address the issue that the optimal solution oracle may have multiple outputs, we consider an additional strongly convex decision regularization function for both the oracle and the ICEO objective. We then prove that the ICEO method is asymptotically consistent and provide finite-sample analysis in the form of generalization bounds. Moreover, we investigate the non-differentiability of the regularized optimal solution oracle which often leads to computational difficulties in calculating the gradients and poor local minima that are hard to escape. We address this issue by approximating the regularized oracle using differentiable functions. We then provide possible approximation error bounds and the corresponding generalization bounds when using the approximated oracle. For the cases when the nominal optimization problem is semi-algebraic, and when the approximated oracle is constructed using polynomial functions, the ICEO problem can be reformulated as a polynomial optimization problem and thus solved for the optimal solution up to arbitrary accuracy by solving a hierarchy of semi-definite problems.

Naturally, there are many potential directions to investigate for future work. One possible direction is to generalize the ICEO framework to the infinite dimensional case where the random parameter is a continuous random variable. Besides, one may also investigate the ICEO framework in the high-dimensional setting or when the data is non-stationary.

Chapter 4

Distributionally Robust Conditional Quantile Prediction with Fixed Design

4.1 Background and Motivation

The objective of quantile prediction is to estimate or predict a quantile of a random variable. This classic yet meaningful topic has numerous applications, including in operations research, economics, and finance. One example is the famous newsvendor problem in the operations research community, where the solution to this problem involves predicting a specific quantile of the random demand (e.g., see [134]). In the sector of finance, the solution to the well-known value-at-risk(VaR) problem involves finding the quantiles of profit and loss (e.g., see [12]). For economic policymakers, two critical indicators, namely, systemic real risk and systemic financial risk, are defined as the quantiles of GDP growth and of financial risk indicator, respectively ([83]). Given that these random variables of interest depend considerably on related covariates/features, *conditional quantile prediction* becomes a factor. Similar to the aim of univariate quantile prediction, that of conditional quantile prediction is to estimate a specific quantile of a particular random variable (i.e., the response variable, say, $y \in \mathbb{R}$) that is conditioned on the information provided by observed related covariates, $x \in \mathbb{R}^p$.

In practice, when historical observations $\{(x_i, y_i)\}_{i=1}^N$ are available, solving the aforementioned conditional quantile prediction problem is inherently tied to the well-known *quantile regression* technique (e.g., see [81, 82, 80]). In a simple linear model setting, the τth conditional quantile of y is assumed to be a linear combination of covariates x, for any $\tau \in (0, 1)$. Moreover, the coefficients are often estimated using the linear quantile regression technique

via solving the following simple optimization problem,

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{N} \rho_\tau(y_i - \beta^T x_i), \tag{4.1}$$

where $\rho_\tau(u) = u(\tau - I_{(u \leq 0)})$ is the quantile loss function (see Sections 1.3 and 1.4 of [81]). It should be noted that solving (4.1) is equivalent (up to scaling) to seeking the optimal policy for the newsvendor problem under the setting in which demand linearly depends on exogenous variables. This problem is also called the *data-driven / big data newsvendor problem* and has been extensively studied thoroughly in existing literature (e.g., [20] and [10]).

To establish theoretical performance bounds, current studies assume the availability of independent and identically distributed (i.i.d.) data, where the i.i.d. assumption is made for both the response variable and the exogenous features ([10], [81]). Although this assumption is appropriate in some limited contexts, it is often unrealistic in many real-life problems. This is because the external features might be predetermined or designed so that the data samples cannot be treated as realizations of a random variable whose law follows an underlying distribution. In these cases, the design points , $x_1, \ldots, x_n$, should be treated as deterministic rather than random. This interpretation of the design points is called the *fixed design* in regression analysis (e.g., [124]). The following are some related examples:

1. Fashion goods demand: Fast fashion brands, such as Zara, H&M, and Gap, issue hundreds of "micro-seasons" per year. This frequent delivery of new products provides opportunities for data-driven solutions in estimating demand and managing inventory. Suppose a fast fashion company aims to forecast sales of a new product based on historical observations of other products. The style, color, and materials often serve as important contextual information for predicting demand. However, these factors are not randomly determined. Instead, they are carefully determined by fashion designers based on certain fashion trends and principles.

2. Promotions and pricing decisions: Promotions are typically used as a price inducement to attract price-conscious buyers who are not interested in regular-priced products. For e-commerce companies, promotions generate more website traffic that can be converted to sales. Therefore, promotions and prices are non-negligible side information when predicting sales. However, promotions and prices are carefully determined, often by the revenue management team according to certain strategies. In fact, promotions and prices are decision variables themselves in many cases.

3. Chemometric problems: Chemometric problems require that statistical methods be applied to chemical data analyses. By producing and analyzing observational data, chemometricians understand and improve industrial chemical processes when conducting food research and environmental studies. The covariates in chemometrics are often

controlled and calibrated based on experimental principles. These settings offer a huge potential for the application of functional regression tools in fixed-design settings ([52]).

4. Pharmacology analysis: In the field of pharmacology, treatment effects are often estimated based on collected observations of amounts of drug dose and outcomes. Thus, it is natural to apply a fixed-design setting to pharmacology problems.

By contrast, let us assume the design points as random fall into the category of random design. The evaluations of out-of-sample performances between fixed- and random-design settings are different. In a random-design setting, an estimator is evaluated on a random set of observations that are generated from the same underlying distribution as the training observations, whereas in the fixed-design setting, an estimator is evaluated on given feature vectors (for details, see [29]).

Adopting a fixed-design setting, we focus on proposing a robust and practical approach for the conditional quantile prediction problem. More specifically, we consider a distributionally robust optimization (DRO) framework. Under this framework, an ambiguity set of possible probability distributions is constructed using available data and/or other prior information, and an objective function is formulated as to minimize the worst-case expected cost over the ambiguity set. We construct the ambiguity set by including all the probability distributions that are sufficiently close to a nominal distribution with respect to the *Wasserstein distance* ([71]). When external features are considered as random with i.i.d. observations, existing results of a data-driven DRO (DD-DRO) approach can be applied. These results use the Wasserstein metric as the distance measure and the empirical distribution as the center/nominal distribution (e.g., [46, 57, 22]). For instance, by leveraging the recent measure concentration results of the empirical distribution (cf. [23, 51]), [46] obtained good out-of-sample performance guarantees. [22] also investigated similar problems under a more general setting and derived theoretical results such as strong duality and estimation of the tolerance region.

However, in contrast to DD-DRO studies that assume the existence of i.i.d. observations, these data are not accessible under the fixed-design setting that we adopt. As the distributions of the samples $\{y_i\}_{i=1}^N$ depend on the exogenous features $\{x_i\}_{i=1}^N$, which are given and not always identically distributed, the $\{y_i\}_{i=1}^N$ are neither. By contrast, setting features as given does not affect the independence of $\{y_i\}_{i=1}^N$.

In this study, we provide a data-driven distributionally robust approach to address the conditional quantile prediction problem. We assume that there are accessible historical observations of predetermined features $\{x_i\}_{i=1}^N$ and of the corresponding response variable $\{y_i\}_{i=1}^N$, which are independent but not identically distributed. The lack of i.i.d. samples makes our framework non-trivially distinguishable from those proposed in existing literature. Thus, it should not be viewed as an easy application of the existing DD-DRO works. More

specifically, we first propose a method that carefully chooses a nominal distribution as the center of a Wasserstein ambiguity set. We then derive the measure concentration results for this nominal distribution under mild conditions. Such measure concentration results further lead to finite-sample performance guarantees. Moreover, in Section 4.6, we show that for the considered quantile prediction problem, the DRO solution with a random-design setting coincides with the sample average approximation (SAA) solution. In other words, the SAA solution maintains a certain level of robustness under the random-design setting. However, in Section 4.7, we demonstrate that for applications that should adopt the fixed-design setting, the SAA solution is no longer robust. Therefore, the proposed DRO approach is necessary for these fixed-design circumstances.

Our contributions can be summarized as follows:

- *DD-DRO with non-i.i.d. observations.* Through the lens of DRO, we examine the quantile prediction problem with observed feature information. In other words, we assume that observations are given in a non-i.i.d. manner, such as through a series of controlled experiments, and that features are observable or designed prior to the decision-making process. To the best of our knowledge, this is the first time that a distributionally robust framework has been built for this type of non-i.i.d. setting.
- *A regress-then-robustify framework with performance guarantees.* We propose a regress-then-robustify framework and consider the ambiguity set as a Wasserstein ball. By choosing a *surrogate empirical distribution* as the center of the Wasserstein ball, we ensure the existence of measure concentration results. Thus, under much milder conditions, we achieve finite-sample performance guarantees and asymptotic consistency.
- *A polynomially solvable reformulation.* Under the aforementioned construction of the Wasserstein ambiguity set, an effective and practical reformulation is achieved. The objective function is decomposed into two simple terms, and the DRO solution is consistent with the empirical quantile solution of ordinary least squares (OLS) residuals, due to the piecewise linear property of the quantile loss. Notably, this reformulation is polynomially solvable using linear regression followed by residual sorting.

The remainder of this paper is organized as follows. Section 4.2 reviews related literature, and Section 4.3 describes details of the problem setting and our two-step DRO framework. In Section 4.4, we state the simple and practical reformulation and in Section 4.5, we develop the finite-sample performance guarantees as well as the asymptotic consistency. A comparison with the conditional quantile prediction problem under a random-design setting is demonstrated in Section 4.6. Section 4.7 reports numerical performances of our proposed DRO method under a fixed-design setting. Section 4.8 discusses possible extensions to the case of heteroskedasticity. Section 4.10 concludes the study.

4.2 Literature Review

Our work extends the data-driven DRO framework to incorporate side information as observed features when i.i.d. samples are assumed to be unavailable. Moreover, we provide a robust data-driven solution to the quantile prediction problem with both finite-sample performance guarantees and asymptotic consistency. Notably, the newsvendor problem, which is one of the most popular and well-studied inventory models, essentially predicts a certain quantile of the demand. Therefore, in this section, we first review the literature on quantile regression and the newsvendor problem when external features are considered and historical data are available. We then review some of the current studies on distributionally robust approaches and end the section with an overview of the studies that considered the non-i.i.d. process under a supervised learning framework.

From the perspective of data-driven decision making, [20] proposed linear programming models to deal with the case in which demand is a linear combination of some exogenous variables and a random shock. [127] then extended this approach to the case in which demand is censored. However, in both studies, the solution involved minimizing an objective of the in-sample cost following the empirical risk minimization idea (also known as the already defined SAA), which essentially assumes that real demand follows the empirical distribution of the samples. Later, [10] incorporated regularization for dimension reduction and proposed a nonparametric approach based on kernel regression. Assuming that demand process is bounded, that the random shock density function has a bounded support, and that samples are i.i.d., the authors derived a high probability bound for a finite-sample optimality gap. Similar to [10], we provide a finite-sample performance bound but through a different DRO framework under milder conditions. Other studies that incorporated features in decision models, including [130] and [60], investigated topics different from ours.

In addition to that described in the aforementioned work ([10]), another method to achieve out-of-sample performance guarantees is to use the concept of *robust optimization*. Its application to newsvendor/quantile prediction problems dates back to 1958, when [129] proposed a min-max solution. Considering all the possible values of the demand, a worst-case solution is used as the decision. Other studies that fall within this stream include [94, 118]. This approach, although it increases robustness, is too conservative in many cases. For example, optimality may be compromised in coping with extreme values of demand that seldom occur.

To address these issues, the DRO framework has gained popularity in recent years. In this setting, an ambiguity set of possible probability distributions of demand is constructed, and the objective function is formulated with respect to the worst-case expected cost over the choice of a distribution in the set. The ambiguity set can be constructed based on either distributional information or historical data.

With certain distributional information, the DRO approach has been applied to the newsvendor/quantile prediction problem. [112] studied the newsvendor problem with partial information about the demand distribution (e.g., mean, variance, symmetry, and unimodality) and derived ordering quantities that minimize the worst-case costs. In special cases, closed-form solutions are derived, and their relationship to entropy maximization is established. Similarly, [155] focused on a case in which the mean of the demand and the variance (or support) are known. Instead of the worst-case newsvendor cost, they attempt to minimize the regret with respect to the expected cost based on complete information. [118] incorporated conditional-VaR-based profit maximization under the assumption of an ellipsoid or box discrete distribution; and [152] demonstrated that optimal ordering decision with discrete demand is quite different from that with continuous demand.

The DD-DRO framework incorporates historical data points into the DRO framework. Various approaches have been developed to construct the ambiguity set. For example, [147] constructed an ambiguity set whereby the observed data achieves a lower-bounded empirical likelihood and [17] used the confidence region of a goodness-of-fit test. Another widely adopted method to construct the ambiguity set is to first determine a nominal distribution based on the observed samples (e.g. empirical distribution), and then include all probability distributions that are close to the nominal distribution, with respect to a certain probabilistic metric. Popular probability metrics include the ϕ-divergence (see [13, 69, 146]), Prohorov metric (see [45]), and Wasserstein distance (see [46, 153, 57, 22, 21]). For example, [119] limited the variance distance between the candidate distributions and the nominal distribution; [57] measured the distribution distance using the Wasserstein distance instead. Among these metrics, the Wasserstein distance stands out because of two important features. The first is its ability to characterize the distance between a discrete and a continuous distribution. The second is the fact that it is a true distance metric that possesses nice properties such as being symmetric with the triangle inequality holds. Thus, the Wasserstein distance has become increasingly popular in recent years and is chosen for this study.

One important application of the aforementioned DD-DRO is to robustify the statistical/supervised learning problems. [1] applied the distributionally robust approach to logistic regression with a Wasserstein ambiguity set. The authors provided a tractable reformulation and encapsulated the classical regularization as special cases. [56] further established the connection between the Wasserstein distributionally robust approaches and regularization. A broad class of loss functions was identified for which the DRO approach with Wasserstein ambiguity set was asymptotically equivalent to a regularization problem with a gradient-norm penalty. [21] also showed that certain machine learning problems with regularization can be viewed as solutions to DRO problems. The authors introduced a method known as robust Wasserstein profile inference for select regularization parameters without cross-validation. In addition to the Wasserstein ambiguity set, [62] applied the f-divergence to classification problems when analyzing a distributionally robust supervised learning problem.

Nevertheless, all of these approaches assume a stationary unchanging environment in which the data points can be considered as i.i.d. samples. This assumption is impractical in many real-world applications. To this end, supervised learning without the i.i.d. assumption has been studied. [137] and [138] analyzed the case of *covariate shift* when the training and test data were from different distributions. [103], [104] and [120] studied a scenario in which observations were drawn from a stationary ϕ-mixing or β-mixing sequence. This type of sequence assumes dependence between observations in which the dependence weakens over time. However, none of these studies that relaxed the assumption of i.i.d. samples considered the robustness of their learning framework. With non-i.i.d. samples, our work is the first to link studies in the machine learning community with those in the robust optimization community.

Based on our results as presented in Section 4.3, the solution of our approach is consistent with that of a currently used procedure in industry in which the quantile of the response variable is estimated by ranking the residuals of a regression model ([34], [151]). This method is easy to implement and has good performance in practice in many different fields. For instance, [42] used the percentiles of residual ranks to estimate childhood obesity, and [50] used the same approach to study the growth of student test scores.

A similar idea to that of our two-step framework has been adopted for conformalized prediction. A conformalized prediction involves first training a regression model on a training set and then using residuals on a validation set to estimate the uncertainty of a future prediction (readers may refer to [132] for a detailed review). [122] and [79] independently investigated the conformalized quantile regression (CQR) method, which combines the method of conformalized prediction and quantile regression. The CQR method provides confidence intervals for quantile estimators with a high probability. [140] then extended the CQR method under the setting of covariate shift. Our work is different from this stream of research in two respects. First, CQR methods have been used to solve for a confidence interval of quantile estimators, whereas we investigate this problem from an optimization viewpoint and provide an upper-bound of the out-of-sample objective cost with high probability. Moreover, the aforementioned works all assumed the training dataset to be exchangeable or i.i.d., whereas we adopt a fixed-design setting in which the exchangeability and i.i.d. assumption do not hold.

4.3 Problem Formulation

In this work, we consider a linear model as follows:

$$y = \beta_0^T x + \varepsilon, \tag{4.2}$$

where $y \in \mathbb{R}$ is the response variable of interest, $x \in \mathbb{R}^p$ is a vector of explanatory variables without an interception term, $\beta_0 \in \mathbb{R}^p$ is the vector of linear coefficients and $\varepsilon \in \mathbb{R}$ is a

random innovation. In addition, ε is assumed to be independent of x and is i.i.d. in all instances with zero mean and unknown variance σ^2. Note that the distribution of ε is not necessarily normal. Since we adopt a fixed-design setting, all feature observations $\{x_i\}_{i=1}^N$ are treated as deterministic rather than random.

Note that the τth quantile of a random variable y is defined as

$$Q_\tau(y) = \inf\{q : \mathbb{P}(y \leq q) \geq \tau\}. \tag{4.3}$$

Under the fixed-design setting that we adopt, we are concerned with out-of-sample performance with some given feature vector $x = c$. By the linear model (4.2) and the independence between x and ε, the conditional quantile admits the following form for any c:

$$\begin{aligned} Q_\tau(y|c) =& Q_\tau(\beta_0^T c + \varepsilon|c) \\ =& \beta_0^T c + Q_\tau(\varepsilon) \\ =& \beta_0^T c + s_\tau \end{aligned} \tag{4.4}$$

where $s_\tau := Q_\tau(\varepsilon)$ is defined as the τth quantile of ε.

In addition to the definition (4.3), $Q_\tau(y)$ can also be defined as the solution of a stochastic optimization problem:

$$Q_\tau(y) = \arg\min_q \mathbb{E}[\rho_\tau(y - q)] \tag{4.5}$$

where ρ_τ is the quantile loss function (also called the check function) which takes the form of

$$\rho_\tau(u) = u(\tau - I_{(u\leq 0)}) = \begin{cases} \tau u & \text{if } u > 0, \\ (\tau - 1)u & \text{if } u \leq 0. \end{cases} \tag{4.6}$$

When no intercept term exists in x, equation (4.4) is identifiable. Thus, β_0 and s_τ are the unique solution to the following stochastic optimization problem, for any fixed value c,

$$\begin{aligned} \beta_0, s_\tau &= \arg\min_{\beta,s} \mathbb{E}_{y|c}[\rho_\tau(y - \beta^T c - s)], \quad \forall c \in \mathbb{R}^p, \\ &= \arg\min_{\beta,s} \mathbb{E}[\rho_\tau(\beta_0^T c + \varepsilon - \beta^T c - s)], \quad \forall c \in \mathbb{R}^p. \end{aligned} \tag{4.7}$$

Note that the expectation is taken over the randomness of ε. The first equation follows when q is replaced by $\beta^T c + s$ in (4.5) and the identifiable property of β_0 and s_τ. The second equation holds because the only randomness of y comes from the noise term ε, and ε does not depend on c.

However, the conditional distribution of $y|c$ is not available because neither the value β_0 nor the distribution of ε is recognized. Under the random-design setting, a common

treatment is to consider the following stochastic optimization problem instead:

$$\min_{\beta,s} \mathbb{E}_{(x,y)}[\rho_\tau(y - \beta^T x - s)]. \tag{4.8}$$

We then apply the SAA approach, where

$$\hat{\beta}_{SAA}, \hat{s}_{SAA} = \arg\min_{\beta,s} \sum_{i=1}^{N} \rho_\tau(y_i - \beta^T x_i - s). \tag{4.9}$$

This is also recognized as quantile regression ([81]). Under the random-design setting, SAA is essentially equivalent to replacing the real joint distribution of (x, y) by its empirical distribution, which converges to the real underlying data generating distribution as the sample size increases.

However, in the fixed-design setting that we explore, this converging empirical distribution property no longer holds, as our data points are not i.i.d. To obtain a robust yet not-too-conservative solution, we consider the DRO approach which hedges against a chosen set of candidate distributions and incorporates the feature information as well:

$$\hat{\beta}_{DRO}, \hat{s}_{DRO} = \arg\min_{\beta,s} \sup_{\mathbb{Q}_y \in \mathcal{P}_y} \mathbb{E}_{\mathbb{Q}_y}[\rho_\tau(y - \beta^T c - s)], \tag{4.10}$$

where $\mathbb{Q}_y$ is a candidate conditional distribution of y given the feature vector c and $\mathcal{P}_y$ is an ambiguity set constructed through observed data points $\{(x_i, y_i)\}_{i=1}^N$. Similarly, $\mathcal{P}_y$ also depends on c, the feature vector of interest. Equivalently, we can rewrite the formulation to the expectation with respect to ε as we assume that it is the source of randomness and is independent of c:

$$\hat{\beta}_{DRO}, \hat{s}_{DRO} = \arg\min_{\beta,s} \sup_{\mathbb{Q}_\varepsilon \in \mathcal{P}_\varepsilon} \mathbb{E}_{\mathbb{Q}_\varepsilon}[\rho_\tau(\beta_0^T c + \varepsilon - \beta^T c - s)], \tag{4.11}$$

where $\mathbb{Q}_\varepsilon$ and $\mathcal{P}_\varepsilon$ denote the candidate distribution and ambiguity set of ε, respectively. As reviewed in Section 4.2, various criteria can be used to construct $\mathcal{P}_\varepsilon$. In this research, we adopt $\mathcal{P}_\varepsilon$ constructed using the Wasserstein distance.

Compared with the traditional random interpretation of the feature vector x, a major difficulty arises with the fixed-design setting. The process ε is the only process that has i.i.d. realizations. However, this process is not explicitly observable. Instead, we can obtain only realizations of $\{y_i\}_{i=1}^N$ where $y_i = \beta_0^T x_i + \varepsilon_i$ and the value of β_0 is also unknown. To decouple the two sources of unknownness, we introduce a new quantity, $\varepsilon(\beta) = \varepsilon + (\beta_0 - \beta)^T c$. Its distribution is the same as that of ε but shifted by a constant $(\beta_0 - \beta)^T c$. Then, the distributionally robust optimization formulation (4.11) can be rewritten as:

$$\hat{\beta}_{DRO}, \hat{s}_{DRO} = \arg\min_{\beta,s} \sup_{\mathbb{Q}_{\varepsilon(\beta)} \in \mathcal{P}(\beta)} \mathbb{E}_{\mathbb{Q}_{\varepsilon(\beta)}}[\rho_\tau(\varepsilon(\beta) - s)], \tag{4.12}$$

where $\mathbb{Q}_{\varepsilon(\beta)}$ and $\mathcal{P}(\beta)$ are the corresponding candidate distribution and ambiguity set of $\varepsilon(\beta)$, respectively. It should be emphasized that our ambiguity set relies not only on the data points and vector c, but also on the value of β, which in turn must be determined. Therefore, with the intention of employing the Wasserstein ambiguity set, both the center and radius of the Wasserstein ball depend on the value of β. Hence, jointly determining the optimal value of β and s cannot be accomplished.

To simplify the problem described in (4.12) and to obtain a tractable solution, we proceed with a *regress-then-robustify* framework by first providing a good estimation of β_0 and then solving an approximated distributionally robust problem. For any estimator of β, we can construct *surrogate samples* of ε with the residuals associated with β, defined as$\{(\varepsilon_i(\beta) = y_i - \beta^T x_i)\}_{i=1}^N$. Then, the ambiguity set is constructed as a Wasserstein ball centered at a discrete distribution with equal mass on the surrogate samples (we call it the *surrogate empirical distribution*) (i.e. $\mathcal{P}(\beta) := \mathbb{B}_{r_N}\big(\hat{\mathbb{P}}_N(\beta)\big)$ with $\hat{\mathbb{P}}_N(\beta) = \frac{1}{N}\sum_{i=1}^N \delta(\varepsilon_i(\beta))$, where $\delta(\varepsilon_i(\beta))$ denotes the unit mass on $\varepsilon_i(\beta)$). In this section, we start by using the least squared estimator $\hat{\beta}_{OLS}^N$ from the classical linear regression as the value for estimating β_0. It is well-known that, under certain conditions, $\hat{\beta}_{OLS}^N$ is a consistent estimator of β_0, and has a nice closed-form solution with well-studied properties which is expressed as:

$$\hat{\beta}_{OLS}^N = (X^T X)^{-1} X^T Y, \tag{4.13}$$

where $X \in \mathbb{R}^{n\times p}$ is the design matrix with x_i being its ith row and $Y \in \mathbb{R}^n$ being a column vector that stores $\{y_i\}_{i=1}^N$.

Let us further define $\varepsilon' = \varepsilon + (\beta_0 - \hat{\beta}_{OLS}^N)^T c$ with unknown distribution $\mathbb{P}_{\varepsilon'}$ (i.e., $\mathbb{P}_\varepsilon$ shifted by $(\beta_0 - \hat{\beta}_{OLS}^N)^T c$). Then, the corresponding surrogate samples are $\{\varepsilon_i^{OLS}\}_{i=1}^N$ with $\varepsilon_i^{OLS} = y_i - (\hat{\beta}_{OLS}^N)^T x_i$ and the surrogate empirical distribution is $\hat{\mathbb{P}}_N(\hat{\beta}_{OLS}^N) = \frac{1}{N}\sum_{i=1}^N \delta(\varepsilon_i^{OLS})$. The surrogate empirical distribution leads to a Wasserstein ball ambiguity set $\mathcal{P}(\hat{\beta}_{OLS}^N) := \mathbb{B}_{r_N}\big(\hat{\mathbb{P}}_N(\hat{\beta}_{OLS}^N)\big)$, where the definition of the Wasserstein distance is provided in Section 4.3.1. Note that ε_i^{OLS} is the linear regression residual of the ith data point, but is not an actual realization of the random variable ε'. Thus, compared with existing studies, the ambiguity set of our formulation is no longer centered at an empirical distribution of a random variable. Instead, it centers at the surrogate empirical distribution based on OLS residuals. Finally, we aim to solve the following distributionally robust optimization problem:

$$\min_{s\geq 0} \sup_{\mathbb{Q}_{\varepsilon'}\in\mathcal{P}(\hat{\beta}_{OLS}^N)} \mathbb{E}_{\mathbb{Q}_{\varepsilon'}}[\rho_\tau(\varepsilon' - s)], \tag{4.14}$$

where $\hat{s}_N$ denotes the optimal solution of (4.14).

4.3.1 Preliminaries

Under our problem setting, we introduce the definition of the Wasserstein distance as follows:

Definition 4.3.1 (Wasserstein distance)**.** Let $\mathbb{P}$ and $\mathbb{Q}$ denote two distributions supported on a set Ξ. Let $\mathcal{M}(\Xi^2)$ denote the set of probability distributions on $\Xi \times \Xi$. The (1-)Wasserstein distance between $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$d_W(\mathbb{P}, \mathbb{Q}) := \inf_{\Pi \in \mathcal{M}(\Xi^2)} \Big\{ \int_{\Xi^2} d(\xi, \xi') \Pi(\mathrm{d}\xi, \mathrm{d}\xi') : \Pi(\mathrm{d}\xi, \Xi) = \mathbb{Q}(\mathrm{d}\xi), \Pi(\Xi, \mathrm{d}\xi') = \mathbb{P}(\mathrm{d}\xi') \Big\}, \tag{4.15}$$

where $d(\xi, \xi')$ is a metric on Ξ.

We remark that the definition stated above is for the 1-Wasserstein distance which is a special case of the a generalized p-Wasserstein distance with $p \geq 1$. In this section, we exclusively focus on the 1-Wasserstein distance as given in Definition 4.3.1. From Section 4.3 to Section 4.5, our analysis focuses on the case in which $\Xi = \mathbb{R}$ and thus admits the absolute difference $|\xi - \xi'|$ as $d(\xi, \xi')$. In addition, we should mention that the Wasserstein metric can also be defined in a dual representation.

Lemma 4.3.1 ([46], Theorem 3.2)**.** *For any distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{M}(\Xi)$ we have*

$$d_W(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{L}} \Big\{ \int_{\Xi} f(\xi) \mathbb{P}(\mathrm{d}\xi) - \int_{\Xi} f(\xi) \mathbb{Q}(\mathrm{d}\xi) \Big\},$$

where $\mathcal{L}$ denotes the space of all Lipschitz functions with $|f(\xi) - f(\xi')| \leq d(\xi, \xi')$ for all $\xi, \xi' \in \Xi$. $d(\xi, \xi')$ denotes the metric on Ξ that is used in Definition 4.3.1.

Because our DRO solution relies on the OLS estimator, its performance depends on how effective $\hat{\beta}^N_{OLS}$ is as a proxy for β_0. Thus, we review the important assumptions for guaranteeing the performance of $\hat{\beta}^N_{OLS}$ together with some properties of the OLS estimator. Throughout our analysis in this study, we also make the same assumptions.

Assumption 5. *The design matrix X is of full rank. This ensures the invertibility of matrix $X^T X$, so that the closed-form solution (4.13) is well-defined.*

Although we proceed with the fixed-design setting in which $\{x_i\}_{i=1}^N$ should not be regarded as i.i.d. samples, we still require the design matrix X to have certain asymptotic properties.

Assumption 6. *As sample size N goes to infinity, the $p \times p$ matrix*

$$M_{xx} = \lim_{N \to \infty} N^{-1} X^T X = \lim_{N \to \infty} N^{-1} \sum_{i=1}^{N} x_i x_i^T \tag{4.16}$$

exists and is finite nonsingular.

This assumption immediately leads to the observation that $c^T(X^TX)^{-1}c \to 0$ as N goes to infinity for any bounded vector c because

$$\begin{aligned}\lim_{N\to\infty} c^T(X^TX)^{-1}c &= \lim_{N\to\infty} \frac{c^T(N^{-1}X^TX)^{-1}c}{N} \\ &= \frac{c^T(\lim_{N\to\infty} N^{-1}X^TX)^{-1}c}{\lim_{N\to\infty} N} \\ &= \frac{c^T M_{xx}^{-1} c}{\lim_{N\to\infty} N} \\ &= 0.\end{aligned} \tag{4.17}$$

Moreover, with the aforementioned two assumptions held, by applying the Markov law of large numbers (Theorem A.9 from [29]), one can prove that $\hat{\beta}_{OLS}^N$ is a (strongly) consistent estimator.

Lemma 4.3.2. *Under Assumptions 5 and 6, the OLS estimator is strongly consistent, i.e.,*

$$\hat{\beta}_{OLS}^N \xrightarrow{a.s.} \beta_0.$$

Closely related to linear regression is the so-called $N \times N$ "hat matrix", that is, the projection matrix defined as

$$H = X(X^TX)^{-1}X^T. \tag{4.18}$$

Its diagonal elements $\{h_{ii}\}_{i=1}^N$ are widely known as the leverage scores of data points, measuring how far away the feature vectors of each observation are from those of the other observations. We state some additional useful properties of the leverage scores in the following lemma.

Lemma 4.3.3 ([32])**.** *Let h_{ii} denote the ith diagonal element of the projection matrix H defined in (4.18), where $X \in \mathbb{R}^{n\times p}$ is the design matrix. Then, the following results hold:*

1. $0 \le h_{ii} \le 1,\ \forall i = 1, 2, \ldots, N.$
2. $trace(H) = \sum_{i=1}^N h_{ii} = p.$

4.4 Tractable Reformulation

In Section 4.3, we proposed a DRO formulation for solving a linear conditional quantile prediction problem. However, formulation (4.14) is a seemly intractable optimization problem with infinite-dimensional constraints. To assure that our formulation is computationally tractable, we further demonstrate that it can be rewritten as a convex program with decoupled components and can thus be solved by OLS regression followed by sorting of the residuals. The following theorem provides a tractable reformulation.

Theorem 4.4.1. *For any $r_N \geq 0$, (4.14) can be reformulated as*

$$\min_{s,\lambda \geq \max\{\tau,1-\tau\}} \lambda r_N + \frac{1}{N}\sum_{i=1}^{N} \rho_\tau(\varepsilon_i^{OLS} - s). \tag{4.19}$$

Proof. Proof of Theorem 4.4.1: By our construction of the ambiguity set, $\mathcal{P}(\hat{\beta}^N_{OLS})) := \mathbb{B}_{r_N}(\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS}))$, for a fixed s, we have

$$\begin{aligned}
&\sup_{\mathbb{Q}_{\varepsilon'} \in \mathbb{B}_{r_N}(\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS}))} \mathbb{E}_{\mathbb{Q}_{\varepsilon'}}[\rho_\tau(\varepsilon' - s)] \\
&= \begin{cases} \sup_{\mathbb{Q}_{\varepsilon'} \in \mathcal{M}(\mathbb{R})} \mathbb{E}_{\mathbb{Q}_{\varepsilon'}}[\rho_\tau(\varepsilon' - s)] \\ s.t. \quad d_W(\mathbb{Q}_{\varepsilon'}, \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) \leq r_N \end{cases} \\
&= \begin{cases} \sup_{\Pi} \int_{\mathbb{R}} \rho_\tau(\varepsilon' - s)\mathbb{Q}_{\varepsilon'}(d\varepsilon') \\ s.t. \int_{\mathbb{R}\times\mathbb{R}} |\varepsilon' - \xi|\Pi(d\varepsilon', d\xi) \leq r_N \\ \text{where } \varepsilon' \sim \mathbb{Q}_{\varepsilon'}, \xi \sim \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS}), \text{and } \varepsilon', \xi \text{ has joint distribution } \Pi, \end{cases}
\end{aligned} \tag{4.20}$$

where the equality (4.20) holds due to the definition of Wasserstein distance (Definition 4.3.1). Let $\mathbb{Q}^i$ denote the conditional distributions of ε' given $\xi = \varepsilon_i^{OLS}$, $\forall i = 1, \ldots, N$ and let $\delta(\varepsilon_i^{OLS})$ denote the Dirac distribution of ξ at $\xi = \varepsilon_i^{OLS}$. Then, for any joint distribution of (ξ, ε'), denoted as Π, we can find $\mathbb{Q}^i$ that satisfy $\Pi = \frac{1}{N}\sum_{i=1}^{N} \delta(\varepsilon_i^{OLS}) \otimes \mathbb{Q}^i$, where $\otimes$ denotes the product of $\delta(\varepsilon_i^{OLS})$, a measure of ξ, and $\mathbb{Q}^i$, a measure of ε'. Thus, (4.20) can be further reformulated as

$$(4.20) = \begin{cases} \sup_{\mathbb{Q}^i \in \mathcal{M}(\mathbb{R})} \frac{1}{N}\sum_{i=1}^{N} \int_{\mathbb{R}} \rho_\tau(\varepsilon' - s)\mathbb{Q}^i(d\varepsilon') \\ s.t. \quad \frac{1}{N}\sum_{i=1}^{N} \int_{\mathbb{R}} \|\varepsilon' - \varepsilon_i^{OLS}\|\mathbb{Q}^i(d\varepsilon') \leq r_N \end{cases} \tag{4.21}$$

$$= \sup_{\mathbb{Q}^i \in \mathcal{M}(\mathbb{R})} \inf_{\lambda \geq 0} \frac{1}{N}\sum_{i=1}^{N} \int_{\mathbb{R}} \rho_\tau(\varepsilon' - s)\mathbb{Q}^i(d\varepsilon') + \lambda(r_N - \frac{1}{N}\sum_{i=1}^{N} \int_{\mathbb{R}} \|\varepsilon' - \varepsilon_i^{OLS}\|\mathbb{Q}^i(d\varepsilon')) \tag{4.22}$$

$$\leq \inf_{\lambda \geq 0} \sup_{\mathbb{Q}^i} \lambda r_N + \frac{1}{N}\sum_{i=1}^{N} \int_{\mathbb{R}} (\rho_\tau(\varepsilon' - s) - \lambda|\varepsilon' - \varepsilon_i^{OLS}|)\mathbb{Q}^i(d\varepsilon'). \tag{4.23}$$

Problem (4.22) follows from (4.21) by adding a dual multiplier λ, and the inequality (4.23) reduces to equality by strong duality which is guaranteed by Theorem 1 of [22].

In addition, the optimal $\mathbb{Q}^i$ that solves

$$\sup_{\mathbb{Q}^i} \frac{1}{N}\sum_{i=1}^{N} \int_{\mathbb{R}} (\rho_\tau(\varepsilon' - s) - \lambda|\varepsilon' - \varepsilon_i^{OLS}|)\mathbb{Q}^i(d\varepsilon')$$

is the Dirac distribution $\mathbb{Q}^i = \delta(\varepsilon_i^*)$, where ε_i^* solves

$$\sup_{\varepsilon' \in \mathbb{R}} \rho_\tau(\varepsilon' - s) - \lambda|\varepsilon' - \varepsilon_i^{OLS}|.$$

Because both $\rho_\tau(\cdot)$ and $|\cdot|$ are piece-wise linear functions, only their slopes, i.e. τ, $\tau - 1$ and λ, play a role in determining the supremum of the function value. It can be easily shown that $\varepsilon_i^* = \varepsilon_i^{OLS}$ if $\lambda \geq \max\{1-\tau, \tau\}$. Otherwise, $\sup_\varepsilon \rho_\tau(\varepsilon' - s) - \lambda|\varepsilon' - \varepsilon_i^{OLS}| = \infty$. Moreover, because this is a minimization problem and we can easily identify a feasible value (e.g. $\lambda = 1$) for the objective value to be finite, the optimal value should be finite. Therefore, we add the constraint $\lambda \geq \max\{\tau, 1-\tau\}$. We then have

$$(4.14) = \min_{s,\lambda \geq 0} \lambda r_N + \frac{1}{N} \sum_{i=1}^N \begin{cases} \rho_\tau(\varepsilon_i^{OLS} - s), & \text{if } \lambda \geq \tau \text{ and } \lambda \geq 1-\tau, \\ \infty, & \text{o.w.} \end{cases} \tag{4.24}$$

$$= \min_{s,\lambda \geq \max\{\tau, 1-\tau\}} \lambda r_N + \frac{1}{N} \sum_{i=1}^N \rho_\tau(\varepsilon_i^{OLS} - s). \tag{4.25}$$

□

The aforementioned theorem provides a convex program reformulation of (4.14) with two decomposed components. This reformulation further leads to a simple closed-form solution where $\hat{\lambda}_N = \max\{\tau, 1-\tau\}$ and $\hat{s}_N = \arg\min_s \frac{1}{N} \sum_{i=1}^N \rho_\tau(\varepsilon_i^{OLS} - s)$. In particular, $\hat{\lambda}_N$ is essentially the Lipschitz constant of ρ_τ, and $\hat{s}_N$ is the τth sample quantile of the OLS residuals $\{\varepsilon_i^{OLS}\}_{i=1}^N$. This solution can be obtained in polynomial time by sorting the linear regression residuals.

It is worth mentioning that sorting the linear regression residuals is a currently used procedure in linear model quantile prediction. In fact, it achieves good practical performances in many different fields including healthcare (see [42]) and education (see [34, 50]). Theorem 4.4.1 thus shows that this procedure has certain robustness and explains its empirical success from the perspective of DRO.

Furthermore, Theorem 4.4.1, combined with Section 4.5, indicates that under a linear fixed design, a simple upper-bound for the out-of-sample quantile loss of our conditional quantile predictor can be calculated by summing the minimal empirical risk of linear regression residuals and a constant term, $\max\{\tau, 1-\tau\} r_N$. This finding aligns with Theorem 10 of [86], which states that for a convex Lipschitz loss function, the worst-case risk with Wasserstein-DRO formulation coincides with Lipschitz-regularized empirical loss.

4.5 Performance Guarantees

Although Theorem 4.4.1 demonstrates that (4.14) has a simple and practical solution, the robustness of the solution still remains to be shown. To do this, in this section, we provide

theoretical performance guarantees for the proposed distributionally robust solution. We first develop measure concentration results for the surrogate empirical distribution. With finite sample points, we provide sufficient conditions for choosing the Wasserstein ball radius such that the out-of-sample quantile loss can be bounded from above with high probability. We then prove the asymptotic optimality of our proposed DRO solution.

4.5.1 Notation and Assumptions

To clarify our analysis, we first introduce the following notation:

– J^*: the optimal expected cost which we target at, i.e.,

$$J^* = \min_{\beta,s} \mathbb{E}_{y|c}[\rho_\tau(y - \beta^T c - s)] = \min_{\beta,s} \mathbb{E}_{\mathbb{P}_{\varepsilon(\beta)}}[\rho_\tau(\varepsilon(\beta) - s)]. \tag{4.26}$$

– $\hat{J}_N$: the in-sample cost achieved by the distributionally robust optimization formulation (4.14), i.e.,

$$\begin{aligned} \hat{J}_N &= \min_{s} \sup_{\mathbb{Q}_{\varepsilon'} \in \mathcal{P}(\hat{\beta}^N_{OLS})} \mathbb{E}_{\mathbb{Q}_{\varepsilon'}}[\rho_\tau(\varepsilon' - s)], \\ \varepsilon' &= \varepsilon + (\beta_0 - \hat{\beta}^N_{OLS})^T c. \end{aligned} \tag{4.27}$$

– J_{oos}: the expected out-of-sample cost achieved by our DRO solution $(\hat{s}_N, \hat{\beta}^N_{OLS})$, i.e.,

$$J_{oos} = \mathbb{E}_{\mathbb{P}_{\varepsilon'}}[\rho_\tau(\varepsilon' - \hat{s}_N)]. \tag{4.28}$$

The feasibility of $(\hat{s}_N, \hat{\beta}^N_{OLS})$ to the original problem (4.11) implies that $J^* \leq J_{oos}$. However, our primary concern is to obtain an upper-bound of J_{oos}. Recall that the DRO formulation minimizes the worst-case in-sample cost $\hat{J}_N$ among the ambiguity set and that both $\hat{J}_N$ and $\mathcal{P}(\hat{\beta}^N_{OLS})$ are random due to the randomness of the data samples. $\hat{J}_N$ bounds J_{oos} from above if and only if the ambiguity set we constructed includes the real distribution $\mathbb{P}_{\varepsilon'}$. Thus, we examine the ambiguity set

$$\begin{aligned} \mathcal{P}(\hat{\beta}^N_{OLS}) &:= \mathbb{B}_{r_N}\left(\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})\right) \\ &:= \left\{\mathbb{Q} \in \mathcal{M}(\mathbb{R}) : d_W(\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS}), \mathbb{Q}) \leq r_N\right\}, \end{aligned} \tag{4.29}$$

where the ambiguity set consists of all distributions within a Wasserstein ball that has radius r_N and center $\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})$. Note that $\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})$ is not constructed with i.i.d. samples. Therefore, the measure concentration results from [51] cannot be directly applied. This means that the performance guarantees demonstrated in existing literature (e.g., [57] and [46]) no longer hold. To reestablish the desired performance guarantees, we follow a "divide and conquer" approach and apply the attractive properties of $\hat{\beta}^N_{OLS}$. Throughout this section, we make the following additional assumption:

Assumption 7 (Light-tailed assumption)**.** *There exists an exponent $a > 1$ such that*

$$A := \mathbb{E}_{\mathbb{P}_\varepsilon}[\exp(\|\varepsilon\|^a)] < \infty.$$

We also assume that $A > 1$.

This light-tailed assumption requires the tail of the distribution of ε to decay at an exponential rate and implies that the variance of ε is finite. In particular, to derive an upper-bound for the variance σ^2, we have the following result:

Lemma 4.5.1. *Under Assumption 7, the variance of ε, denoted by σ^2 can be upper-bounded by light-tail coefficients:*

$$\sigma^2 \leq (2(A-1))^{\frac{1}{a}}.$$

Proof. Proof of Lemma 4.5.1: Since $|\varepsilon| \geq 0$, by Taylor's expansion, we have

$$A := \mathbb{E}_{\mathbb{P}_\varepsilon}[\exp(|\varepsilon|^a)] \geq \mathbb{E}_{\mathbb{P}_\varepsilon}[1 + |\varepsilon|^a + \frac{1}{2}|\varepsilon|^{2a}] \tag{4.30}$$

$$\geq 1 + \frac{1}{2}\mathbb{E}_{\mathbb{P}_\varepsilon}[|\varepsilon|^{2a}] \tag{4.31}$$

$$\geq 1 + \frac{1}{2}\sigma^{2a} \tag{4.32}$$

where the second inequality is a consequence of $|\varepsilon| \geq 0$ as well. The last inequality holds due to Jensen's inequality. □

Thus, we can assume that we know a constant κ such that $\sigma^2 \leq \kappa$.

4.5.2 Finite-sample Performance Guarantees

In our problem setting, the measure concentration results of [51] no longer hold because the ambiguity set is centered at the surrogate empirical distribution. Therefore, we first investigate the measure concentration of the surrogate empirical distribution and then use it as a building block for establishing finite-sample performance guarantees.

Theorem 4.5.2 (Measure Concentration)**.** *Suppose Assumptions 5, 6 and 7 hold, then for any $\eta \in (0,1)$, there exists r_N such that $r_N \to 0$ as $N \to \infty$ and*

$$\mathbb{P}^N\{d_W(\mathbb{P}_{\varepsilon'}, \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) \geq r_N\} \leq \eta,$$

where $\mathbb{P}^N(\cdot)$ denotes the joint distribution of the N samples of ε.

Corollary 4.5.2.1 (Finite-sample Performance Guarantee)**.** *It follows immediately from Theorem 4.5.2 that $\hat{J}_N$ upper-bounds J_{oos} with high probability. i.e.,*

$$\mathbb{P}^N(J_{oos} \leq \hat{J}_N) \geq 1 - \eta.$$

We first introduce the following lemmas to prove Theorem 4.5.2 using a "divide and conquer" approach.

When β takes its real value β_0, $\hat{\mathbb{P}}_N(\beta_0)$ boils down to an empirical distribution of ε based on i.i.d. samples. Thus, the result from [51] holds for $\hat{\mathbb{P}}_N(\beta_0)$.

Lemma 4.5.3 ([51] and [46]). *If Assumption 7 holds, then for any $\eta' \in (0,1)$, we can set $r_{1,N}$ such that*

$$\mathbb{P}_\varepsilon^N\{d_W(\mathbb{P}_\varepsilon, \hat{\mathbb{P}}_N(\beta_0)) \geq r_{1,N}\} \leq \eta'$$

for all $N \geq 1$, where

$$r_{1,N}(\eta') := \begin{cases} (\frac{\log(c_1\eta'^{-1})}{c_2 N})^{\frac{1}{2}} & \text{if } N \geq \frac{log(c_1\eta'^{-1})}{c_2} \\ (\frac{\log(c_1\eta'^{-1})}{c_2 N})^{\frac{1}{a}} & \text{if } N < \frac{log(c_1\eta'^{-1})}{c_2} \end{cases} \tag{4.33}$$

where c_1, c_2 are positive constants that only depend on a and A (see Theorem 2 of [51] for details regarding choice of c_1 and c_2). Note that $r_{1,N}$ chosen according to (4.33) converges to 0 as $N \to \infty$.

We further investigate the Wasserstein distance between the real empirical distribution $\hat{\mathbb{P}}_N(\beta_0)$ and its approximation by the surrogate empirical distribution $\hat{\mathbb{P}}_N(\hat{\beta}_{OLS}^N)$.

Lemma 4.5.4. *If Assumptions 5, 6 and 7 hold, then by setting $r_{2,N} = \sqrt{\frac{p\kappa}{N}}\frac{1}{\eta'}$, we have*

$$\mathbb{P}_\varepsilon\{d_W(\hat{\mathbb{P}}_N(\beta_0), \hat{\mathbb{P}}_N(\hat{\beta}_{OLS}^N)) \geq r_{2,N}\} \leq \eta' \tag{4.34}$$

for all $\eta' \in (0,1)$, and that $r_{2,N} \to 0$ as $N \to \infty$.

Proof. Proof of Lemma 4.5.4: Let ε_i^{OLS} denote the linear regression residual of the ith data point, then

$$\begin{aligned} d_W(\mathbb{P}_N(\beta_0), \mathbb{P}_N(\hat{\beta}_{OLS}^N)) =& \sup_{f\in\mathcal{L}}\{\frac{1}{N}\sum_{i=1}^N f(\varepsilon_i) - f(\varepsilon_i^{OLS})\} \\ \leq& \frac{1}{N}\sum_{i=1}^N |\varepsilon_i - \varepsilon_i^{OLS}| \\ =& \frac{1}{N}\sum_{i=1}^N |(\hat{\beta}_N^{OLS} - \beta_0)^T x_i| \\ =& \frac{1}{N}\sum_{i=1}^N |x_i^T(X^TX)^{-1}X^T\boldsymbol{\varepsilon}|, \end{aligned} \tag{4.35}$$

where the first equality follows from the dual representation of the Wasserstein metric (Lemma 4.3.1) and the following inequality is justified by f being a 1-Lipschitz function. We then stack the realizations of random noises, $\{\varepsilon_i\}_{i=1}^N$, in vector $\boldsymbol{\varepsilon}$. To simplify the notation, let $V_i = x_i^T(X^TX)^{-1}X^T\boldsymbol{\varepsilon}$. Then, with X being fixed and independent of $\boldsymbol{\varepsilon}$, we have

$$\mathbb{E}(V_i) = x_i^T(X^TX)^{-1}X^T\mathbb{E}(\boldsymbol{\varepsilon}) = 0$$

and

$$\begin{aligned}\mathrm{Var}(V_i) &= \sigma^2 x_i^T(X^TX)^{-1}x_i \\ &\leq \kappa h_{ii},\end{aligned}$$

where κ is an upper-bound of σ^2 as defined in Lemma 4.5.1 and $h_{ii} = x_i^T(X^TX)^{-1}x_i$ is the leverage of the ith data point in linear regression as defined in lemma 4.3.3. It follows that (4.35) can be bounded from below as

$$\mathbb{P}(\frac{1}{N}\sum_{i=1}^N |V_i| > r_{2,N}) \leq \frac{\mathbb{E}[\frac{1}{N}\sum_{i=1}^N |V_i|]}{r_{2,N}} \tag{4.36}$$

$$\leq \frac{\frac{1}{N}\sum_{i=1}^N \sqrt{\mathbb{E}V_i^2}}{r_{2,N}} \tag{4.37}$$

$$= \frac{\frac{1}{N}\sum_{i=1}^N \sqrt{\mathrm{Var}(V_i)}}{r_{2,N}} \tag{4.38}$$

$$\leq \frac{\sqrt{\frac{1}{N}\sum_{i=1}^N \mathrm{Var}(V_i)}}{r_{2,N}} \tag{4.39}$$

$$\leq \frac{\sqrt{\frac{1}{N}\sum_{i=1}^N \kappa h_{ii}}}{r_{2,N}}$$

$$\leq \sqrt{\frac{\kappa p}{N}}\frac{1}{r_{2,N}}. \tag{4.40}$$

The inequality (4.36) follows from Markov's inequality. (4.37) and (4.39) both hold according to Jensen's inequality. The equality (4.38) is a consequence of the fact that $\mathbb{E}[V_i] = 0$. Finally, when Lemma 4.3.3 is applied, (4.40) holds. We then have

$$\begin{aligned}\mathbb{P}_\varepsilon\{d_W(\hat{\mathbb{P}}_N(\beta_0), \hat{\mathbb{P}}_N(\hat{\beta}_{OLS}^N)) \geq r_{2,N}\} &\leq \mathbb{P}(\frac{1}{N}\sum_{i=1}^N |V_i| > r_{2,N}) \\ &\leq \sqrt{\frac{\kappa p}{N}}\frac{1}{r_{2,N}}.\end{aligned}$$

Thus, with $r_{2,N} = \sqrt{\frac{p\kappa}{N}}\frac{1}{\eta'}$, we can bound the distance, $d_W(\hat{\mathbb{P}}_N(\beta_0), \hat{\mathbb{P}}_N(\hat{\beta}_{OLS}^N))$, with any desired probability $(1-\eta') \in (0,1)$. Because both p and κ are finite constant, for any given value of η', we have $r_{2,N} \to 0$ as $N \to \infty$. □

Finally, it remains to bound the distance between the underlying true distributions of ε and $\varepsilon' = \varepsilon + (\beta_0 - \hat{\beta}^N_{OLS})^T c$, where the latter is simply the former shifted by a constant. We then have the following lemma:

Lemma 4.5.5. *If Assumption 7 holds, then for any $\eta' \in (0,1)$, we can set*

$$r_{3,N} = \sqrt{\frac{\kappa c^T (X^T X)^{-1} c}{\eta'}}$$

so that

$$\mathbb{P}_\varepsilon\{d_W(\mathbb{P}_\varepsilon, \mathbb{P}_{\varepsilon'}) \geq r_{3,N}\} \leq \eta'. \tag{4.41}$$

Note that $r_{3,N} \to 0$ as $N \to \infty$ by (4.17).

Proof. Proof of Lemma 4.5.5: Noting the fact that given the OLS estimator, $\hat{\beta}^N_{OLS}$, ε' is simply the random variable ε shifted by a constant. We show that a general case of Lemma 4.5.5, as stated in Lemma 4.5.6, holds.

Lemma 4.5.6. *Let μ and ν denote two probability measures supported on $\mathbb{R}$. Suppose $d\nu(x+t) = d\mu(x)$ for all $x \in \mathbb{R}$ with a constant t. Then it holds that*

$$d_W(\mu, \nu) = |t|$$

Proof. Proof of Lemma 4.5.6: Because of the symmetry of this problem, without loss of generality, we may assume $t \geq 0$. For any joint density function $h(x,y)$ that has marginal $\mu(x)$ and $\nu(y)$, if $d\nu(x+t) = d\mu(x)$, then we have

$$\int_{\mathbb{R}} h(y, x+t)dy = \int_{\mathbb{R}} h(x,y)dy := g(x).$$

Thus,

$$\begin{aligned}
\inf_{h\in\mathcal{M}(\mathbb{R}^2)} \int_{\mathbb{R}\times\mathbb{R}} |x-y| h(x,y)dxdy &\geq \inf_{h\in\mathcal{M}(\mathbb{R}^2)} \int_{\mathbb{R}\times\mathbb{R}} (y-x)h(x,y)dxdy && (4.42)\\
&= \inf_{h\in\mathcal{M}(\mathbb{R}^2)} \Big(\int_{\mathbb{R}} yh(x,y)dxdy - \int_{\mathbb{R}} xh(x,y)dxdy\Big) && (4.43)\\
&= \inf_{g\in\mathcal{M}(\mathbb{R})} \Big(\int_{\mathbb{R}} yg(y-t)dy - \int_{\mathbb{R}} xg(x)dx\Big) && (4.44)\\
&= \inf_{g\in\mathcal{M}(\mathbb{R})} \Big(\int_{\mathbb{R}} (y+t)g(y)dy - \int_{\mathbb{R}} yg(y)dy\Big) && (4.45)\\
&= t && (4.46)
\end{aligned}$$

Here, we slightly abuse the notation and let $\mathcal{M}(\mathbb{R}^2)$ denote the set of probability density functions of all probability distributions on $\mathbb{R} \times \mathbb{R}$. Similarly, we let $\mathcal{M}(\mathbb{R})$ denote the set of density functions of all marginal distributions on $\mathbb{R}$.

We then show that the aforementioned inequality reduces to equality by specifying a joint distribution that achieves the lower-bound. Consider $h(x, y) = \delta(y = x + t)d\mu(x)/dx$, where δ denotes the delta function. Using the fundamental property of the δ function, $\int_{\mathbb{R}} h(x, y)dy = 1_{y=x+t}d\mu(x)/dx$. Thus, with this specific $h(x, y)$,

$$\int_{\mathbb{R}\times\mathbb{R}} |x - y|h(x, y)dxdy = \int_{\mathbb{R}} td\mu(x) = t.$$

□

Thus, by applying Lemma 4.5.6 to the left-hand-side of (4.41), we obtain the Wasserstein distance between $\mathbb{P}_\varepsilon$ and $\mathbb{P}_{\varepsilon'}$ as

$$d_W(\mathbb{P}_\varepsilon, \mathbb{P}_{\varepsilon'}) = |(\beta_0 - \beta^N_{OLS})^T c|.$$

Then, by Chebyshev's Inequality, we have

$$\begin{aligned}
\mathbb{P}^N(|(\beta_0 - \hat{\beta}^N_{OLS})^T c| > r_{3,N}) &\leq \frac{\mathrm{Var}(|(\beta_0 - \hat{\beta}^N_{OLS})^T c)|)}{(r_{3,N})^2} \\
&\leq \frac{\mathrm{Var}((\beta_0 - \hat{\beta}^N_{OLS})^T c))}{(r_{3,N})^2} \\
&= \frac{\mathrm{Var}(((X^T X)^{-1} X^T \boldsymbol{\varepsilon})^T c)}{(r_{3,N})^2} \\
&= \frac{\sigma^2 c^T (X^T X)^{-1} c}{(r_{3,N})^2} \\
&\leq \frac{\kappa}{(r_{3,N})^2} (c^T (X^T X)^{-1} c).
\end{aligned}$$

It thus suffices to select $r_{3,N} = \sqrt{\frac{\kappa c^T (X^T X)^{-1} c}{\eta'}}$ so that $\mathbb{P}_\varepsilon\{d_W(\mathbb{P}_\varepsilon, \mathbb{P}_{\varepsilon'}) \geq r_{3,N}\}$ is bounded by η'. With Assumption 7, η' converges to zero as N goes to infinity. □

Based on previously stated lemmas, we then prove the measure concentration results for $\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})$.

Proof. Proof of Theorem 4.5.2: By triangle inequality, the Wasserstein distance between $\mathbb{P}_{\varepsilon'}$, which is the underlying true distribution of ε', and $\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})$, which is the surrogate

empirical distribution, can be decomposed into the following three components

$$d_W(\mathbb{P}_{\varepsilon'}, \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) \leq d_W(\mathbb{P}_{\varepsilon}, \hat{\mathbb{P}}_N(\beta_0)) + d_W(\hat{\mathbb{P}}_N(\beta_0), \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) + d_W(\mathbb{P}_{\varepsilon'}, \mathbb{P}_{\varepsilon}), \tag{4.47}$$

where the three components are bounded by Lemma 4.5.3, 4.5.4 and 4.5.5, respectively, with certain probability. In particular, for any $\eta \in (0,1)$, we can first select $\eta' = \frac{\eta}{3}$ and then select $r_{1,N}$, $r_{2,N}$, and $r_{3,N}$ according to the aforementioned Lemmas. Thus, by setting $r_N = r_{1,N} + r_{2,N} + r_{3,N}$, we have

$$\begin{aligned}
\mathbb{P}^N(\mathbb{P}_{\varepsilon'} \in \mathbb{B}_{r_N}(\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS}))) =& \mathbb{P}^N(d_W(\mathbb{P}_{\varepsilon'}, \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) \leq r_N) \\
\geq& 1 - \mathbb{P}(d_W(\mathbb{P}_{\varepsilon}, \hat{\mathbb{P}}_N(\beta_0)) > r_{1,N}) - \mathbb{P}(d_W(\hat{\mathbb{P}}_N(\beta_0), \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) > r_{2,N}) \\
& - \mathbb{P}(d_W(\mathbb{P}_{\varepsilon'}, \mathbb{P}_{\varepsilon}) > r_{3,N}) \\
\geq& 1 - 3\eta' \\
=& 1 - \eta.
\end{aligned}$$

□

4.5.3 Asymptotic Consistency

Section 4.5 demonstrated that, with properly chosen r_N, $\hat{J}_N$ is an upper-bound of the expected out-of-sample cost J_{oos} with any desired probability with finite data points. It still remains to explore the limiting behavior of this upper-bound when the sample size tends to infinity. Hereafter, we study the asymptotic behavior of the DRO solution and show that the gap between $\hat{J}_N$ and J^* asymptotically converges to zero.

Theorem 4.5.7 (Asymptotic Consistency)**.** *Suppose that Assumptions 5, 6 and 7 hold and that a sequence $\eta_N \in (0,1)$ satisfies $\sum_{N=1}^{\infty} \eta_N < \infty$ and $\lim_{N\to\infty} r_N(\eta_N) = 0$ (e.g., a possible choice is $\eta_N = \exp\{-\sqrt{N}\}$). Then*

a) $\hat{J}_N \downarrow J^*$ as $N \to \infty$ where J^* is the optimal value of (4.26);

b) Any accumulation point of $\{(\hat{s}_N, \hat{\beta}^N_{OLS})\}$ is $\mathbb{P}^\infty_\varepsilon$-almost surely one of the optimal solution for (4.26).

Proof. Proof of Theorem 4.5.7: We first show that, with a sequence η_N described in the theorem statement, $\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS}), N \in \mathbb{N}$, converges under Wasserstein metric to $\mathbb{P}_{\varepsilon'}$ almost surely with respect to $\mathbb{P}^\infty_\varepsilon$. Theorem 4.5.2 implies that $\mathbb{P}^N\{d_W(\mathbb{P}_{\varepsilon'}, \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) \leq r_N(\eta_N)\} \geq 1 - \eta_N$. As $\sum_{N=1}^{\infty} \eta_N < \infty$, then the Borel–Cantelli Lemma ([77], Theorem 2.18) can be applied and we have

$$\mathbb{P}^\infty_\varepsilon\{d_W(\mathbb{P}_{\varepsilon'}, \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) \leq r_N(\eta_N) \text{ for all sufficiently large } N\} = 1. \tag{4.48}$$

Then since $\lim_{N\to\infty} r_N(\eta_N) = 0$, $\lim_{N\to\infty} d_W(\mathbb{P}_{\varepsilon'}, \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) = 0$ almost surely.

As the DRO solution is feasible for the original stochastic program (4.26), we always have $J^* \leq J_{oos}$. In addition, for any sequence of $\eta_N \in (0,1)$, we can choose $r_N(\eta_N)$ based on the conditions derived from Theorem 4.5.2 and Lemma 4.5.3-4.5.5 such that

$$\mathbb{P}^N\{J^* \leq \hat{J}_N\} \geq \mathbb{P}^N\{J_{oos} \leq \hat{J}_N\} \geq 1 - \eta_N, \ \forall N \in \mathbb{N} \tag{4.49}$$

Then we apply the Borel–Cantelli Lemma and have

$$\mathbb{P}^\infty_\varepsilon\{J^* \leq \hat{J}_N, \ \forall \text{ sufficiently large } N\} = 1,$$

and

$$\mathbb{P}^\infty_\varepsilon\{J_{oos} \leq \hat{J}_N, \ \forall \text{ sufficiently large } N\} = 1. \tag{4.50}$$

Next, it still remains to show that

$$\mathbb{P}^\infty_\varepsilon\{\limsup_{N\to\infty} \hat{J}_N \leq J^*\} = 1. \tag{4.51}$$

According to Theorem 4.4.1, recall that s_τ is the τ-th quantile of ε,

$$\begin{aligned}
\limsup_{N\to\infty} \hat{J}_N =& \limsup_{N\to\infty} r_N \max\{\tau, 1-\tau\} + \min_s \mathbb{E}_{\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})}[\rho_\tau(\varepsilon - s)] \\
\leq& \limsup_{N\to\infty} r_N \max\{\tau, 1-\tau\} + \mathbb{E}_{\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})}[\rho_\tau(\varepsilon - s_\tau)] \qquad (4.52)\\
=& \limsup_{N\to\infty} r_N \max\{\tau, 1-\tau\} + \mathbb{E}_{\mathbb{P}_\varepsilon}[\rho_\tau(\varepsilon - s_\tau)] + \mathbb{E}_{\hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})}[\rho_\tau(\varepsilon - s_\tau)] - \mathbb{E}_{\mathbb{P}_\varepsilon}[\rho_\tau(\varepsilon - s_\tau)] \\
=& \limsup_{N\to\infty} r_N \max\{\tau, 1-\tau\} + \mathbb{E}_{\mathbb{P}_\varepsilon}(\rho_\tau(\varepsilon - s_\tau)) + d_W(\mathbb{P}_\varepsilon, \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) \qquad (4.53)\\
\leq& \limsup_{N\to\infty} r_N \max\{\tau, 1-\tau\} + \mathbb{E}_{\mathbb{P}_\varepsilon}(\rho_\tau(\varepsilon - s_\tau)) + d_W(\mathbb{P}_\varepsilon, \mathbb{P}_{\varepsilon'}) + d_W(\mathbb{P}_{\varepsilon'}, \hat{\mathbb{P}}_N(\hat{\beta}^N_{OLS})) \\
& \qquad (4.54)\\
=& \mathbb{E}_{\mathbb{P}_\varepsilon}(\rho_\tau(\varepsilon - s_\tau)) \quad \mathbb{P}^\infty_\varepsilon\text{-almost surely} \qquad (4.55)\\
=& J^*,
\end{aligned}$$

where (4.52) follows from the feasibility of s_τ, (4.53) holds due to the dual representation of Wasserstein distance as defined in Lemma 4.3.1 and the fact that $\rho_\tau(\cdot)$ is 1-Lipschitz. Inequality (4.54) follows from the triangle inequality of Wasserstein distance. Moreover, we further notice that $d_W(\mathbb{P}_\varepsilon, \mathbb{P}_{\varepsilon'}) \to 0$, $\mathbb{P}^\infty_\varepsilon$-almost surely, due to the almost surely convergence of $\hat{\beta}^N_{OLS}$. Then according to (4.48) and the fact that $r_N \to 0$, (4.55) holds. Finally, the claim $\limsup_{N\to\infty} \hat{J}_N \leq J^*$ then leads to the conclusion that $\hat{J}_N \downarrow J^*$.

We next want to show that any limit point of $(\hat{s}_N, \hat{\beta}^N_{OLS})$, if exists, is an optimal solution of J^*. In fact, as a quantile of a random variable minimizes its expected quantile loss, we

know that s_τ, β_0 is the minimizer to (4.26) and that it is unique when the quantile function is identifiable. Because of the strong consistency of $\hat{\beta}^N_{OLS}$ from Lemma 4.3.2, i.e. $\hat{\beta}^N_{OLS} \to \beta_0$, $\mathbb{P}^\infty_\varepsilon$-almost surely, it remains to show that $\bar{s}$, any limit point of $\hat{s}_N$, if exists, is s_τ.

Since J^* is the optimal objective value and $\bar{s}$ is a feasible solution, we have

$$\begin{aligned} J^* &\leq \mathbb{E}_{\mathbb{P}_\varepsilon}[\rho_\tau(\varepsilon - \bar{s})] \\ &= \mathbb{E}_{\mathbb{P}_\varepsilon}[\lim_{N\to\infty} \rho_\tau(\varepsilon - \hat{s}_N)] && (4.56) \\ &\leq \liminf_{N\to\infty} \mathbb{E}_{\mathbb{P}_\varepsilon}[\rho_\tau(\varepsilon - \hat{s}_N)] && (4.57) \\ &= \liminf_{N\to\infty} \mathbb{E}_{\mathbb{P}_{\varepsilon'}}[\rho_\tau(\varepsilon' - \hat{s}_N)] && (4.58) \\ &\leq \lim_{N\to\infty} \hat{J}_N \\ &= J^*, \end{aligned}$$

where (4.56) follows because $\bar{s}$ is an accumulation point and ρ_τ is a continuous function. Moreover, (4.57) follows from the Fatou's Lemma and (4.58) holds due to the almost surely convergence of $\hat{\beta}^N_{OLS}$ to β_0. □

4.6 Comparison with the Random-design Setting

In this section, we investigate the direct application of distributionally robust optimization using the Wasserstein metric in solving newsvendor problem under the random-design setting. When covariates are considered as random, we let ζ denote the joint random variable (x, y), and $\{\zeta_i\}_{i=1}^N := \{(x_i, y_i)\}_{i=1}^N$ are i.i.d. samples. This section describes a direct application of the methods developed by [46] to address the conditional quantile prediction problem with random-design setting.

The ambiguity set is centered at the empirical distribution $\hat{\mathbb{P}}_N := \frac{1}{N}\sum_{i=1}^N \delta(\zeta_i)$ and it follows that the DRO formulation would be

$$\min_\beta \sup_{\mathbb{Q}\in\mathbb{B}_{r_N}(\hat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\rho_\tau(y - \beta^T x)], \tag{4.59}$$

where r_N is determined by Theorem 3.4 in [46], and certain performance guarantees and asymptotic consistency directly follow. Recall that the Wasserstein distance (4.15) integrates a metric $d(\cdot,\cdot)$ on the support Ξ. Under the random-design setting, (x, y) is the multivariate random variable of interest and $\Xi = \mathbb{R}^p \times \mathbb{R}$. Thus, selecting different distance metrics leads to different reformulations.

We want to focus on the case in which L_1-norm is selected as the distance matrix. The following proposition shows that, under this case, the reformulation also admits a neat and simple form, where the optimal solution coincides with the SAA solution.

Proposition 4.6.0.1. *If* $\|((x,y),(x',y')\|_1 := \sum_{j=1}^{p}|x_j - x'_j| + |y - y'|$ *is considered as the metric on* $\mathbb{R}^p \times \mathbb{R}$, *(4.59) can be reformulated as*

$$\min_{\beta} r_N \max\{\tau, 1-\tau\} + \frac{1}{N}\sum_{i=1}^{N}\rho_\tau(y_i - \beta^T x_i). \tag{4.60}$$

Similar to Theorem 4.4.1, the aforementioned reformulation also consists of decoupled terms, because of the piece-wise linearity of both the L_1-norm and $\rho_\tau(\cdot)$. Moreover, the second term $\frac{1}{N}\sum_{i=1}^{N}\rho_\tau(y_i - \beta^T x_i)$ coincides with the objective function of the SAA approach. Therefore, the solution of (4.60) is essentially the same as the SAA solution, and an upper-bound of the optimal objective value can be derived by adding the term $r_N \max\{\tau, 1-\tau\}$ to the SAA objective value.

This result derives from the Lipschitz property of $\rho_\tau(\cdot)$ and aligns with Theorem 10 of [86]. In Section 4.7, we investigate the numerical performance of the solution of (4.60) (equivalent to the SAA solution) under a circumstance in which the fixed-design setting is more suitable. The numerical results show that, although the solution of (4.60) is robust under the random-design setting, its numerical performance is very unstable and therefore no longer robust in cases in which the fixed-design setting should be adopted.

Proof. Proof of Proposition 4.6.0.1: Let $\zeta := (x,y)$ and $l_\tau(\zeta) := \rho_\tau(y - \beta^T x)$,

$$\sup_{\mathbb{Q}\in\mathbb{B}_{r_N}(\hat{\mathbb{P}}_N)} \mathbb{E}_Q[\rho_\tau(y-\beta^T x)] \tag{4.61}$$

$$= \begin{cases} \sup_{\Pi,\mathbb{Q}} \int_\Xi l_\tau(\zeta)\mathbb{Q}(d\zeta) \\ s.t. \int_{\Xi\times\Xi}\|\zeta - \zeta'\|_1 \Pi(d\zeta, d\zeta') \le r_N \end{cases} \tag{4.62}$$

$$= \begin{cases} \sup_{\mathbb{Q}^i} \frac{1}{N}\sum_{i=1}^{N}\int_\Xi l_\tau(\zeta)\mathbb{Q}^i \\ s.t. \frac{1}{N}\sum_{i=1}^{N}\int_\Xi \|\zeta - \zeta'\|_1 \mathbb{Q}^i \le r_N \end{cases} \tag{4.63}$$

$$= \inf_{\lambda\ge 0}\sup_{Q^i} \lambda r_N + \frac{1}{N}\sum_{i=1}^{N}\int_\Xi (l_\tau(\zeta) - \lambda\|\zeta - \zeta_i\|_1)\mathbb{Q}^i(d\zeta) \tag{4.64}$$

$$= \inf_{\lambda} r_N\lambda + \frac{1}{N}\sum_{i=1}^{N}\sup_{\zeta}(l_\tau(\zeta) - \lambda\|\zeta - \zeta_i\|_1) \tag{4.65}$$

$$= \inf_{\lambda} r_N\lambda + \frac{1}{N}\sum_{i=1}^{N} l_\tau(\zeta_i) \tag{4.66}$$

The last equality holds since ζ solves $\sup_\zeta(l_\tau(\zeta) - \lambda\|\zeta - \zeta_i\|_1)$, and the other equalities follow from the proof of Theorem 4.2 in [46]. □

4.7 Numerical Experiments

In this section, we validate the theoretical results by conducting numerical experiments using a dataset that is simulated based on real-world data. Our objective is to forecast aggregated bike demand during a certain hour of the day based on external information such as temperature and wind speed. An accurate prediction of the quantiles of the random demand process will facilitate better inventory management of bikes. Section 4.7.1 describes the data generating process and verifies the fixed-design setting using hypothesis testing. Section 4.7.2 gives a thorough performance comparison between the proposed DRO and the standard SAA methods. Finally, Section 4.7.3 compares the DRO method with several data-driven prescriptive methods proposed by [18] and [10].

4.7.1 Data

The data used in our experiments are simulated based on a real dataset obtained from [48]. By selecting all observations recorded on weekdays from their training set, we obtain a dataset with $7,412$ instances. We further preprocess the dataset by extracting features that include hour of the day and day of the week. Subsequently, dummy variables are created for the categorical features. Some features are removed to eliminate the multicollinearity issue. Thus, we eventually have a design matrix X with 28 features.

Feature	Type	Quantity	Description
weather	Binary	2	Dummy variables for three different types of weather
temperature	Numeric	1	Temperature
humidity	Numeric	1	Humidity
windspeed	Numeric	1	Wind speed
hour	Binary	21	Dummy variables indicating the hour of the day[1]
wday	Binary	1	Dummy variable indicating if the data is observed on weekdays
timeperiod	numeric	1	Time period index

[1] Some hours are consolidated together as they do not show significance in the fitted linear model.

Table 4.1: A Summary of Features

Table 4.1 summarizes the information of the features in the design matrix. Since the observations correspond to periods of time, the observations should not be viewed as i.i.d., and we should consider the fixed-design setting. Table 4.1 shows that most of the features are time series or are generated based on some trend over time. Therefore, the observations should not be viewed as i.i.d. and we should apply the fixed-design setting. Firstly, we conduct a linear regression model on the aforementioned preprocessed dataset. We then discard the demand column of the dataset and only keep the columns of features. After

that, we construct a partial-synthetic data set by reconstructing the demand column using simulation. More specifically, in simulation, we first take the estimated coefficients as the coefficients for the underlying true linear relationship (β_0). Then demand observations are generated by adding i.i.d. simulated noise, that is,

$$d_i = \beta_0^T x_i + \varepsilon_i. \tag{4.67}$$

Regarding the time-series nature of the dataset, we divide the observations into training and test sets based on time periods. More specifically, we select observations from the most recent 500 time periods for testing. Next, for any particular size N, the training set consists of N observations, counting backwards from the earliest test set observation. Figure 4.1 visualizes our train–test splitting method.

Figure 4.1: Train-test Splitting Method

Typically, the dummy variables for temporal events, namely, *hour*, *wday*, and *timeperiod* are not generated by a random process. In addition, according to the time-series nature of our data, the remaining feature variables, namely *weather*, *temperature*, *humidity*, and *windspeed*, should be closely correlated within days close to one another and should not be viewed as i.i.d. samples from a random distribution. This dataset thus serves as a practical example of when the random-design assumption fails, and the fixed-design setting is more appropriate.

To validate further the non-i.i.d. nature of this dataset, we use hypothesis tests to check whether any of *weather*, *temperature*, *humidity*, or *windspeed* can be treated as independently generated from a random distribution. To be more specific, we perform a one-sample runs test on training observations for each of the aforementioned features. The null hypothesis of this test is that the values are independent, random observations (we refer readers to Section 4.5 of [135] for a detailed description of the one-sample runs test). We then perform a two-sample (Wald–Wolfowitz) runs test for each of the aforementioned features on both training and test observations. The null hypothesis of this test is that the two samples come from the same unspecified distribution (cf. Section 6.5 of [135]). All of the resulting p-values take the value < 0.0001 which indicate that for each of the aforementioned features, the training and test observations do not originate from the same distribution, and the training observations themselves are not from some underlying random distribution.

We additionally emphasize that, for our numerical experiments as well as any other cases in which the observations are time-related and split into training and test sets by time, the random-design setting is no longer valid because the observations cannot be viewed as randomly generated from an underlying distribution. Therefore, in these cases, it is more appropriate to adopt the fixed-design setting, which does not include any randomness assumptions of the design matrix.

4.7.2 Comparison with SAA

In this part, we evaluate the performance of the DRO method and compare it with the SAA method. Using the SAA method as a benchmark algorithm has a twofold objective: it conforms with its use as a widely adopted method in existing literature ([127, 10]) and it represents the solution of standard Wasserstein DRO method under the random-design setting as demonstrated in Section 4.6. In addition, since the distribution of the noises and β_0 are known in the experiments, we can evaluate the optimal solutions.

For each run of the experiments, with a given size of the training set, namely, N, the values of the response variable in training set are generated by applying (4.67) to the training feature vectors. Thus, N feature vectors exist that correspond to N data points in the training set. The DRO and SAA methods are trained using training set. For each of the 500 feature vectors in the test set, i.i.d. noises are simulated 50 times, resulting in $500 \times 50 = 25,000$ data points in the test set. In other words, each of the 500 different features vectors in the test set results in 50 data points when different simulated noises are added. The out-of-sample performances of the optimal solution, the SAA and DRO methods are evaluated by the average cost on the test set, denoted by J_oos_opt, J_oos_saa, and J_oos_dro, respectively.

Figures 4.2 and 4.3 visualize the out-of-sample performances of different methods under different configurations of N, σ, and the quantile τ. The subplots indicate that the proposed DRO approach outperforms the SAA method under all scenarios, particularly when the sample size is small. As sample size increases, both the DRO and SAA methods converge to the optimal solution. Moreover, the SAA method exhibit unstable performances and is dominated by the DRO method, particularly when a small sample size is employed. Since the data points in the training and test sets are not randomly generated in an i.i.d. manner, the performance guarantees no longer hold for either the SAA method (e.g., [10]) or the DRO method under the random design (e.g., [86]). Therefore, it is not surprising that the SAA benchmark (equivalently, the random-design DRO) is unstable under this setting. Table 4.2 shows the quantified out-of-sample performances. In this table, the percentage values stated in the parentheses correspond to the improvements of our DRO strategy over the SAA approach, which are calculated as $\frac{J\_oos\_dro - J\_oos\_saa}{J\_oos\_saa}$. These values indicate that our proposed DRO policy achieves better quantile predictions.

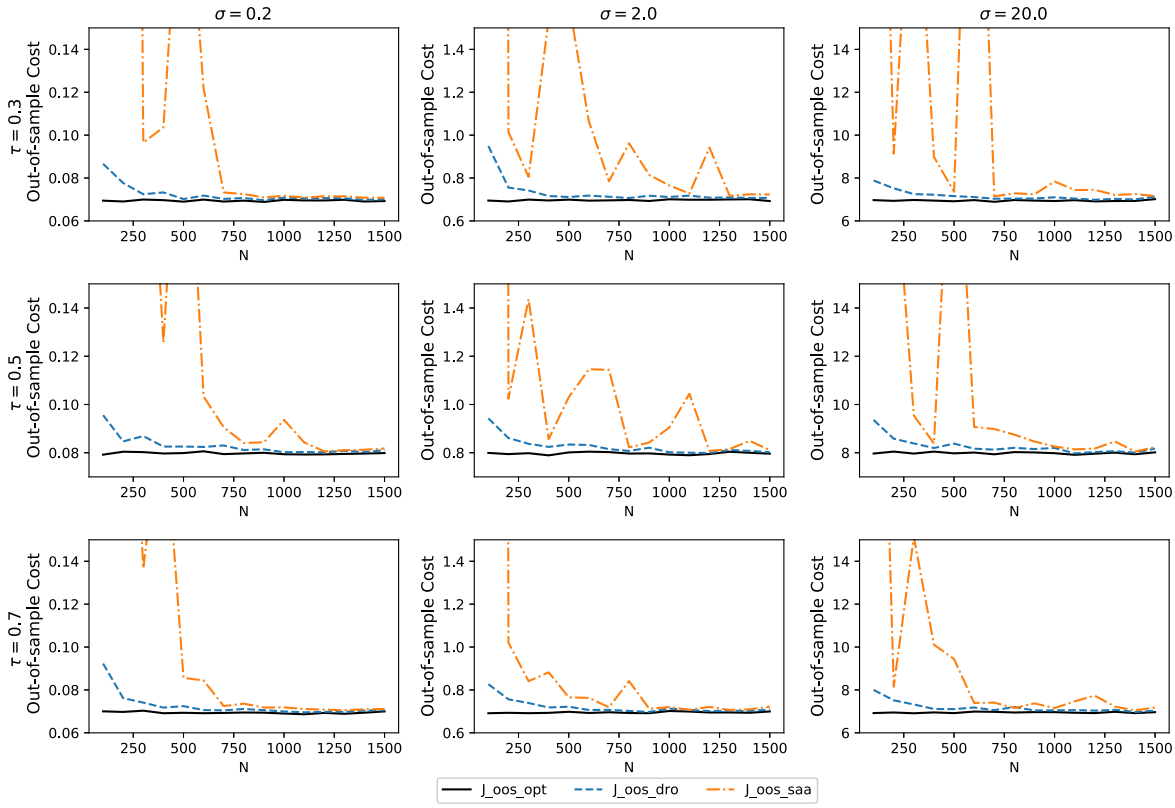


Figure 4.2: Comparison of optimal, SAA and DRO approaches with Gaussian noises.

4.7.3 Comparison with Non-parametric Prescriptive Methods

In the second part of the numerical experiments, we compare the proposed DRO method with two additional benchmark algorithms proposed in [18] and [10]. We employ the non-parametric method motivated by k-nearest-neighbors regression (we refer to Section 2.1 in [18] for details of this method) and the method motivated by Nadaraya-Watson kernel regression (we refer to Section 2.2 in [18] and Section 2.3.2 in [10]) for details of this method). For the Nadaraya-Watson kernel regression method, we use the Gaussian kernel in the experiments. The two aforementioned benchmarks are denoted by KNN and KR, respectively, in the remaining part of this section.

Figures 4.4 and 4.5 visualize the out-of-sample performances of the optimal solution, and solutions by SAA, DRO, KR and KNN methods. Table 4.3 summarizes the corresponding out-of-sample costs of these methods. The hyper-parameters of KR and KNN methods, namely the bandwidth of KR and the number of clusters of KNN, are tuned by enumerating over a grid and selecting the values associated with best test-set performances. We should mention that, although the tuning of hyper-parameters is different from the standard cross-

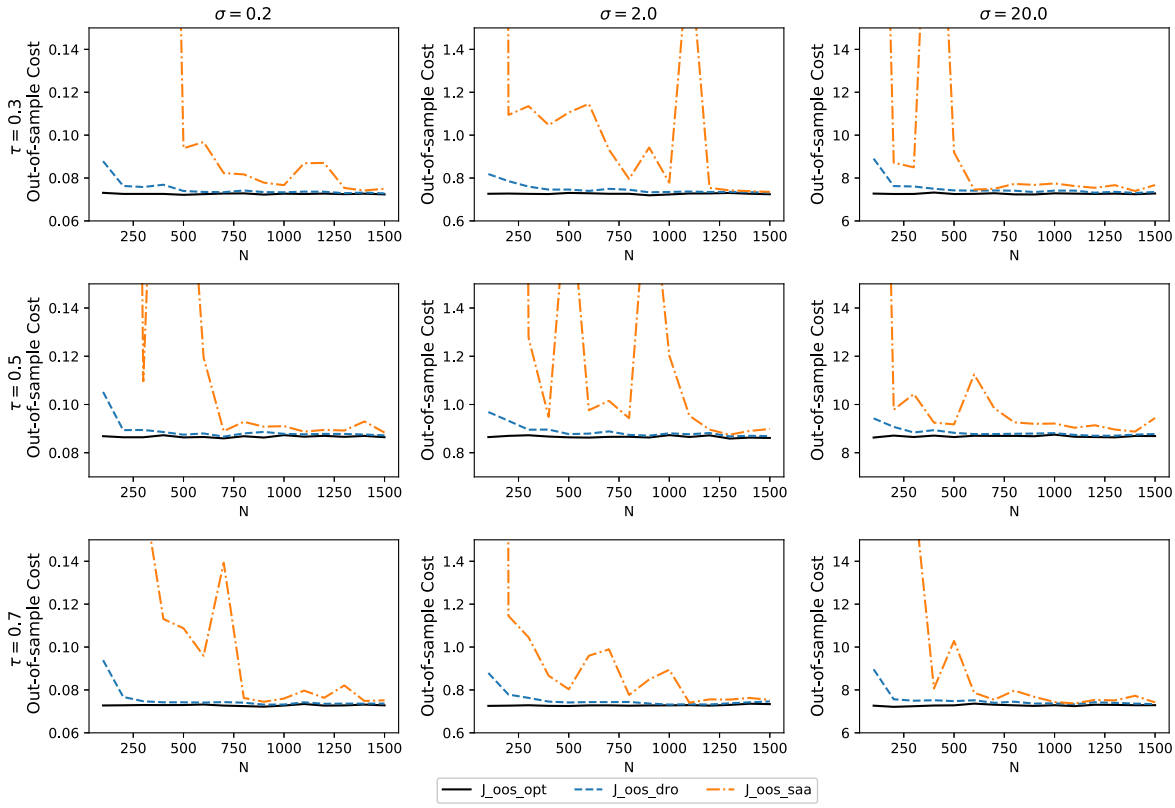


Figure 4.3: Comparison of optimal, SAA and DRO approaches with Uniform noises.

validation approaches, it consistently yields a better out-of-sample result than that of cross-validation, and is thus biased in favor of the benchmark methods (KNN and KR).

In both Figures 4.4 and 4.5, $J_{oos\_kr\_1}$ and $J_{oos\_kr\_2}$ correspond to the best two test-set performances achieved by the KR method from among all candidate bandwidths. Similarly, $J_{oos\_knn\_1}$ and $J_{oos\_knn\_2}$ correspond to two best two test-set results achieved during the tuning procedure employed for KNN. Table 4.3 reports the out-of-sample performances of the optimal solution, SAA, our proposed DRO, as well as the best test-set performances of KR and KNN from among all candidate hyper-parameters.

Table 4.3 and Figures 4.4 and 4.5 demonstrate the advantages of the proposed DRO method. Compared with the SAA method, which is also asymptotically consistent, the DRO method is observed to be more stable and performs better, particularly when sample size is small. Moreover, even though our reporting criterion is biased in favor of the KR and KNN methods, the DRO method consistently outperforms these benchmarks regardless of the sample size.

Noise	σ	N	OPT			SAA			DRO		
			$\tau = 0.3$	0.5	0.7	$\tau = 0.3$	0.5	0.7	$\tau = 0.3$	0.5	0.7
Gaussian	0.2	100	0.069	0.079	0.070	38.27	64.87	40.88	0.087 (−99.8%)	0.096 (−99.9%)	0.092 (−99.8%)
		500	0.069	0.080	0.069	0.247	0.261	0.086	0.070 (−71.7%)	0.083 (−68.2%)	0.072 (−16.3%)
		1500	0.069	0.080	0.070	0.071	0.082	0.071	0.070 (−1.4%)	0.081 (−1.2%)	0.071 (−0.0%)
	2.0	100	0.695	0.799	0.691	42.79	55.88	40.57	0.949 (−97.8%)	0.943 (−98.3%)	0.827 (−98.0%)
		500	0.699	0.801	0.698	1.621	1.030	0.766	0.711 (−56.1%)	0.834 (−19.0%)	0.722 (−5.7%)
		1500	0.694	0.796	0.699	0.723	0.810	0.721	0.707 (−2.2%)	0.802 (−1.0%)	0.708 (−1.8%)
	20.0	100	6.970	7.967	6.921	44.33	59.29	44.796	18.89 (−57.4%)	9.358 (−84.3%)	7.999 (−82.1%)
		500	6.917	7.975	6.920	7.361	9.06	8.495	7.183 (−2.4%)	8.169 (−9.8%)	7.106 (−16.4%)
		1500	7.016	8.016	6.987	8.231	8.138	7.183	7.068 (−14.1%)	8.171 (+0.4%)	7.033 (−2.1%)
Uniform	0.2	100	0.073	0.087	0.072	43.63	54.71	41.66	0.088 (−99.8%)	0.105 (−99.8%)	0.084 (−99.8%)
		500	0.072	0.086	0.073	0.093	0.242	0.078	0.074 (−20.4%)	0.087 (−64.0%)	0.075 (−3.8%)
		1500	0.072	0.086	0.073	0.075	0.088	0.075	0.073 (−2.7%)	0.087 (−1.1%)	0.074 (−1.3%)
	2.0	100	0.729	0.865	0.726	47.23	53.56	42.17	0.864 (−98.2%)	0.969 (−98.2%)	0.879 (−97.9%)
		500	0.724	0.864	0.725	0.950	1.960	0.803	0.743 (−21.8%)	0.877 (−55.3%)	0.742 (−7.6%)
		1500	0.725	0.861	0.734	0.737	0.898	0.753	0.730 (−0.9%)	0.869 (−3.2%)	0.746 (−0.9%)
	20.0	100	7.275	8.631	7.240	46.27	48.58	42.97	8.906 (−80.8%)	9.424 (−80.6%)	8.124 (−81.1%)
		500	7.274	8.651	7.230	9.19	9.176	10.30	7.424 (−19.2%)	8.827 (−3.8%)	7.539 (−26.8%)
		1500	7.279	8.692	7.276	7.67	9.439	7.791	7.350 (−4.2%)	8.776 (−7.02%)	7.379 (−5.3%)

Table 4.2: Comparison of optimal, SAA and DRO approaches

When comparing the performances of SAA, KR, and KNN, we note some additional findings: the KR and KNN are more stable than the SAA method. This is one of the advantages of the nonparametric prescriptive methods because in these methods, more weights are placed on training samples that are close to test set feature vectors. However, both KR and KNN have much slower convergence rates than those of SAA and DRO, and we fail to observe the convergence to the optimal solution with the given data. Therefore, in general, when the sample size is small, KR and KNN may outperform the SAA method; however, with a larger sample size, the SAA method seems to perform better. Based on our training-test splitting method, the later observations in the training set are more likely to be similar to vectors in the test set. Since both the KR and KNN method emphasize training data points that have similar feature vectors as the test set features, as N increases (training data includes samples from earlier time periods), these methods may not be able to gain much useful information from additional training samples. As our currently used tuning method already favors the benchmark algorithms, the previously mentioned findings will still hold should the hyper-parameters be tuned with cross-validation.

4.8 Extension to Heteroskedasticity

In this section, we discuss possible extensions to heteroskedastic linear models in which the i.i.d. assumption of noises $\varepsilon_i s$ is relaxed. Under the setting of heteroskedastic models, the variance of ε_i varies and may depend on the feature vector x_i. Let $\Sigma := \mathbb{E}((y - \mathbb{E}y)(y - \mathbb{E}y)^T)$

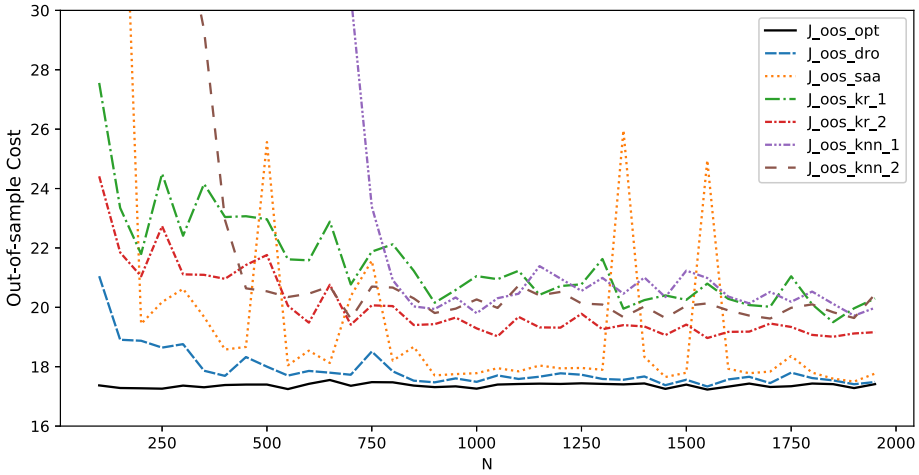


Figure 4.4: Out-of-Sample Comparison with Gaussian noises, $\tau = 0.3$ and $\sigma = 50$.

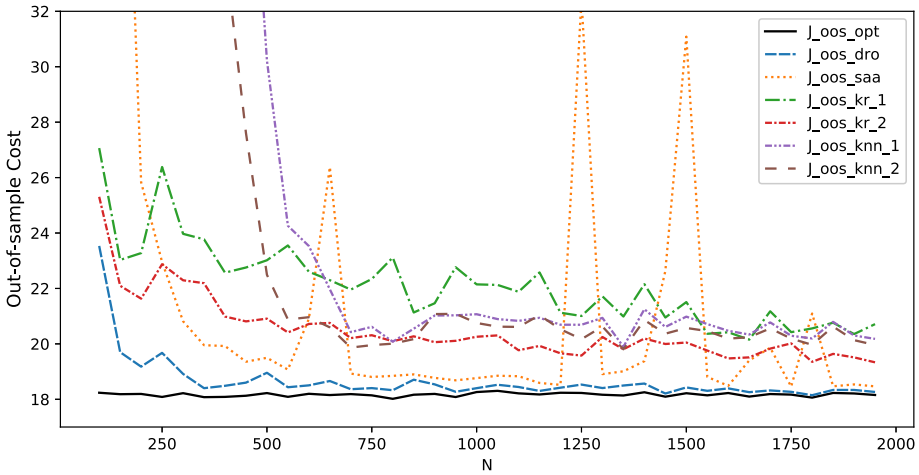


Figure 4.5: Out-of-Sample Comparison with Uniform noises, $\tau = 0.3$ and $\sigma = 50$.

denote the covariance matrix of $\{y_i\}_{i=1}^N$, which may depend on covariates $\{x_i\}_{i=1}^N$. We suppose that with the given feature vector c, the variance of test set noise $\varepsilon(c)$ takes the value of σ^2. We can then rewrite $\Sigma = \sigma^2\Omega$ with a positive semi-definite matrix Ω.

Our approach can be extended to the case in which the matrix Ω is known and the test set noise $\varepsilon(c)$ is independent of training set noises $\{\varepsilon_i\}_{i=1}^N$. Under such a setting, we let W denote the square root matrix of Ω^{-1} (i.e. $\Omega^{-1} = W^T W$). Then, instead of using X and Y as described in Section 4.3.1, we consider $\tilde{X} = WX$ and $\tilde{Y} = WY$. Moreover, the OLS

Noise	σ	N	OPT		SAA		DRO		KR		KNN	
			$\tau = 0.3$	0.7	$\tau = 0.3$	0.7	$\tau = 0.3$	0.7	$\tau = 0.3$	0.7	$\tau = 0.3$	0.7
Gaussian	5.0	100	1.74	1.74	41.42	10.54	2.13	2.08	9.64	13.67	23.20	11.76
		500	1.73	1.73	1.84	2.32	1.78	1.85	7.19	6.67	6.73	7.42
		1500	1.72	1.74	1.81	1.80	1.75	1.76	5.80	6.95	6.36	5.94
	50.0	100	17.37	18.13	48.69	36.78	21.04	20.54	24.41	25.92	63.44	68.65
		500	17.39	18.17	25.56	19.26	18.00	18.49	21.76	21.32	20.53	20.91
		1500	17.39	18.16	17.79	18.74	17.56	18.28	19.41	20.39	20.05	20.84
Uniform	5.0	100	1.82	1.81	44.15	29.80	2.24	2.10	21.98	12.38	53.31	41.48
		500	1.81	1.82	1.97	2.01	1.85	1.85	7.17	6.61	7.18	6.56
		1500	1.81	1.82	1.88	1.90	1.84	1.84	6.58	6.87	5.69	5.69
	50.0	100	18.23	18.13	53.16	36.79	23.53	20.54	25.31	25.93	58.51	68.65
		500	18.22	18.18	19.49	19.26	18.96	18.49	20.91	21.32	22.49	20.76
		1500	18.21	18.16	18.81	19.72	18.30	18.56	19.75	20.35	20.46	20.26

Table 4.3: Comparison of optimal, SAA, DRO and Nonparametric Benchmarks

estimator $\hat{\beta}^N_{OLS}$ should be replaced with the generalized least squares estimator (GLS)

$$\hat{\beta}^N_{GLS} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y} = (X^T\Omega^{-1}X^T)^{-1}X^T\Omega^{-1}Y.$$

The proposed framework can then be directly applied to $\tilde{X}$ and $\tilde{Y}$ if Assumptions 5 and 6 hold with respect to $\tilde{X}$ and $\tilde{Y}$.

One example of the aforementioned case is the multiplicative heteroskedasticity model. Under the setting of the multiplicative heteroskedasticity model in which $y_i = \beta_0^T x_i + u_i$ and the noise u_i also depends on the corresponding covariate x_i. More specifically, according to a given function $h(x_i, \theta)$, $u_i = h(x_i, \theta)\varepsilon_i$ with i.i.d. $\varepsilon_i, i = 1, \ldots, N$ and parameter θ. When the function $h(\cdot, \theta)$ is known, both Ω and $\sigma^2 := h(c, \theta)Var(\varepsilon_i)$ are known, our proposed method can be applied accordingly.

If the distribution of noise has a richer dependence on the covariates, for example, when the test set noise depends on the training set noises, our approach is no longer applicable. This is because $Q_\tau(y)$, the τth quantile of y, depends not only on covariate c (as stated in (4)–(6)), but also on all noises realized in training set $\{\varepsilon_i\}_{i=1}^n$. Therefore, under these circumstances, the problem requires a different formulation/approach.

4.9 Extension to a General Lipschitz Objective Function

The finite-sample performance guarantees provided in Theorem 4.5.2 and Corollary 4.5.2.1 can be directly extended to the case where the objective function ρ_τ is replaced by any ob-

jective function ρ that is 1−Lipschitz. Herein, we slightly abuse the notation and redefine J^*, $\hat{J}_N$, and J_{oos} as:

- J^*: the optimal expected cost which we target at, i.e.,

$$J^* = \min_{\beta,s} \mathbb{E}_{y|c}[\rho(y - \beta^T c, s)] = \min_{\beta,s} \mathbb{E}_{\mathbb{P}_{\varepsilon(\beta)}}[\rho(\varepsilon(\beta), s)].$$

- $\hat{J}_N$: the in-sample cost achieved by the distributionally robust optimization formulation, i.e.,

$$\hat{J}_N = \min_{s} \sup_{\mathbb{Q}_{\varepsilon'} \in \mathcal{P}(\hat{\beta}^N_{OLS})} \mathbb{E}_{\mathbb{Q}_{\varepsilon'}}[\rho(\varepsilon', s)],$$
$$\varepsilon' = \varepsilon + (\beta_0 - \hat{\beta}^N_{OLS})^T c.$$

- J_{oos}: the expected out-of-sample cost achieved by our DRO solution $(\hat{s}_N, \hat{\beta}^N_{OLS})$, i.e.,

$$J_{oos} = \mathbb{E}_{\mathbb{P}_{\varepsilon'}}[\rho(\varepsilon', \hat{s}_N)].$$

The asymptotic consistency can also be easily extended to this case. Herein, we describe how the proof can be adjusted to accommodate any 1-Lipschitz objective function ρ. Note that Equations (4.49)-(4.50) still hold. In the following two lemmas, we prove that $\limsup_{N\to\infty} \hat{J}_N \leq J^*$ when ρ_τ is replaced by ρ.

Lemma 4.9.1. *With any 1-Lipschitz objective function ρ, $\limsup_{N\to\infty} \hat{J}_N \leq J^*$.*

Proof. Proof of Lemma 4.9.1: Because ρ is 1-Lipschitz continuous in ε, we are able to choose any $\delta > 0$ with a δ-optimal decision s_δ such that

$$\mathbb{E}_{\mathbb{P}_{\varepsilon(\beta)}}[\rho(\varepsilon(\beta), s_\delta)] \leq J^* + \delta,$$

and a δ-optimal distribution $\mathbb{Q}_N \in \mathcal{P}(\hat{\beta}^N_{OLS})$ such that

$$\sup_{\mathbb{Q}_{\varepsilon'} \in \mathcal{P}(\hat{\beta}^N_{OLS})} \mathbb{E}_{\mathbb{Q}_{\varepsilon'}}[\rho(\varepsilon', s_\delta)] \leq \mathbb{E}_{\mathbb{Q}_N}[\rho(\varepsilon', s_\delta)] + \delta.$$

Then, we have

$$\begin{aligned}
\limsup_{N\to\infty} \hat{J}_N &\leq \limsup_{N\to\infty} \sup_{\mathbb{Q}_{\varepsilon'} \in \mathcal{P}(\hat{\beta}^N_{OLS})} \mathbb{E}_{\mathbb{Q}_{\varepsilon'}}[\rho(\varepsilon', s_\delta)] \\
&\leq \limsup_{N\to\infty} \mathbb{E}_{\mathbb{Q}_N}[\rho(\varepsilon', s_\delta)] + \delta \\
&\leq \limsup_{N\to\infty} \mathbb{E}_{\mathbb{P}_{\varepsilon(\beta)}}[\rho(\varepsilon', s_\delta)] + d_W(\mathbb{P}_{\varepsilon(\beta)}, \mathbb{Q}_N) + \delta \\
&= \mathbb{E}_{\mathbb{P}_{\varepsilon(\beta)}}[\rho(\varepsilon', s_\delta)] + \delta = J^* + 2\delta.
\end{aligned}$$

The third inequality holds because of the dual representation as defined in Lemma 4.3.1 and the first equality holds due to the convergence of distributions as stated in (4.48). As δ is an arbitrary positive constant, we have $\limsup_{N\to\infty} \hat{J}_N \leq J^*$. □

With Lemma 4.9.1, the asymptotic consistency result in Theorem 4.5.7 can be easily re-established with any 1-Lipschitz objective function ρ.

4.10 Conclusion

This work studies the quantile prediction problem under the setting in which the response variable is a linear function of some covariate variables plus an i.i.d. random noise. In addition, the values of the covariates are given before predictions are made.

We explore the problem in a data-driven environment with the goal of developing a DRO solution. In contrast to the random-design setting widely adopted in the literature, we proceed with a fixed-design setting. We demonstrate that this fixed-design setting is a better fit for many real-life applications because features may be pre-designed or temporaly correlated thus should not be treated as i.i.d. samples. Then, by leveraging the OLS estimators of linear coefficients, we develop a regress-then-robustify distributionally robust framework when samples are no longer i.i.d.

By deriving new concentration inequalities, we bound the expected out-of-sample cost with any desired probability and demonstrate that the solution is asymptotically optimal as the sample size increases. In addition, we show that the proposed distributionally robust framework achieves a neat and practical reformulation that can be solved in polynomial time. We also discuss the equivalence of the SAA approach and existing Wasserstein DRO approaches under the random-design setting.

We conduct numerical experiments to evaluate the performance of the proposed DRO method and then compare it with the SAA method (equivalently, the DRO method with random design) and two nonparametric prescriptive methods. The results reveal that the proposed DRO solution outmatches the benchmarks under a fixed-design setting. We also discuss possible extensions of our proposed method when the i.i.d. assumption of noises is relaxed.

Chapter 5

Learning Newsvendor Problems with Intertemporal Dependence and Moderate Non-stationarities

5.1 Background and Motivation

The newsvendor problem is one of the most classical models in inventory management. To solve the newsvendor problem it often relies on the characterization of the unknown demand in a specific time period. We investigate the newsvendor problem in the case where there are also contextual information available on which the distribution of the demand depends. To characterize or predict the stochastic demand in each period, the contextual information may include both static information, such as product-level information (brands and category), and sequentially updated information, such as weather, historical demands, economic or popularity indicators, and observable strategies of a competitor.

In a big-data environment where rich and high-quality observations of demand and contextual information are available, classic procedures of solving the newsvendor problem adopts two steps: first, predicting demand given the contextual information, and subsequently, solving for the optimal decision based on the predicted demand. This two-step framework, despite being illustrative, may lose information owing to the decoupling of the two steps. Recently, an alternative approach has been established for solving data-driven contextual optimization problem. Such approaches combine the two steps by learning a *data-to-decision mapping* that involves directly mapping data (contextual information) to decisions. Because the contextual information may vary in different time periods, an effective data-to-decision mapping allows a decision-maker to use any contextual information as input and directly obtain a well-performed decision. Such a data-to-decision mapping is usually constructed using machine learning techniques together with a common assumption that observed data are independent and identically distributed (i.i.d.). In most cases, this

i.i.d. assumption is highly restrictive in real-world situations, because historical data are generated and recorded sequentially in a time-ordered approach, instead of being collected from independent sources or experiments. Therefore, temporal dependencies frequently arise from the sequential nature of data. Moreover, historical observations may also present moderate nonstationarities across different periods. For instance, weather and product popularity frequently play serve as contextual information, both exhibit a certain trend and seasonality. This work considers such comparatively more realistic scenarios where intertemporal dependence and moderate non-stationrities exist.

Reliable out-of-sample performance guarantees are crucial for such data-to-decision solutions. This is because the out-of-sample performance guarantees measure the performance of the effective decisions generated by machine learning algorithms in a future contextual period. For this purpose, one key performance guarantee is to control the *generalization error* ([91, 90, 10, 18, 9]). More specifically, the generalization error quantifies the difference between the performance of a learning algorithm evaluated on training samples and on unseen data. Because the performance on a training sample is typically known, the generalization error measures how good an algorithm performs with the outcomes in the future. We aim to develop *generalization bounds*, which provide the bounds of the generalization error with a desired probability. Classic generalization bounds are developed with one (or both) of the assumptions: (i) the observed data are i.i.d.; and (ii) the demand process is bounded with a known upper-bound. However, in most cases of real operations, the demand process is neither i.i.d. nor bounded. Even under the cases in which uncertainties are bounded, it can be difficult for decision-makers to evaluate the exact value of the bound without being loose. To address these issues, we analyze the generalization error of the learned data-to-decision mapping under relatively more realistic settings that includes intertemporal correlations and unbounded randomness.

In this study, we investigate using machine learning methods to learn the data-to-decision mappings for the newsvendor problem, and develop out-of-sample performance guarantees when the standard assumptions of i.i.d. and boundedness are relaxed. Our technical contributions can be summarized as the following:

1. First, we relax the independent assumption and develop out-of-sample performance guarantees when data generated in real operations are stationary with intertemporal dependences. Specifically, we consider scenarios referred to as *mixing*, in which such an intertemporal dependence between two time periods drops as the distance between these two time periods increases. It is aligned with the intuition that for the characterization of the contextual uncertainties in the future, more recent historical data are generally more informative.

2. We provide performance guarantees in the form of generalization bound. Compared to the existing generalization bound provided in [104] for bounded mixing process,

we generalize the existing result to the scenario of unbounded data. To achieve this, we propose the concept of truncated coupling samples while the corresponding performance guarantee depends on the selection of truncated threshold. This concept and techniques also can to be applied to a general case other than newsvendor problem.

3. Moreover, we relax both the independent assumption and the identically distributed assumption and consider moderate nonstationarities. We develop out-of-sample generalization bounds for scenarios in which the underlying data sequence can be viewed as a nonhomogenenous Markov process with moderate non-stationarities in the evolution dynamics.

The remainder of this paper is organized as follows. Section 5.2 reviews the related literature and Section 5.3 describes the learning framework for the contextual newsvendor problem. In Section 5.4, we state the out-of-sample performance of the solution learned from data generated by a stationary, intertemporally dependent process. In Section 5.5, we consider non-stationary processes and provide the performance bound using the data generated from nonhomogeneous Markov processes. In section 5.6, the numerical performances demonstrate the benefits of including intertemporally dependent features. Section 5.7 summarize our findings and concludes this paper.

5.2 Contribution to Literature

Data-driven newsvendor problem with performance guarantees. Newsvendor problem is one of the most popular and well-studied inventory models. The study of data-driven newsvendor problem dates back to [20] who investigated the newsvendor problem in the case where demand is a linear combination of contextual variables and a random shock, without investigating any out-of-sample performance. [90, 91] then analyze the newsvendor problem and propose a sample average approximation (SAA) approach under the assumption of i.i.d. demand. [91] show that the SAA method is approximately optimal (with respect to the relative regret) with a high probability. [90] further tighten the bound proposed in [91]. Both [91] and [90] provide high-probability bounds for a relative out-of-sample cost under the setting of nonconsideration of contextual information. The most closely related literature to our work is [10], which use SAA with regularization and nonparametric methods to learn the data-to-decision mapping for newsvendor problem. While the performance guarantees provided by [10] can only be applied to the case when data is i.i.d., our results are applicable to the case with intertemporal dependence and nonstationarities. Moreover, we also relax the assumption that the demand is upper-bounded.

Data-driven solutions with performance guarantees have also been developed for contextual decision making without contextual information. [36] related the estimation and optimization using a technique called "contextual statistics" for the newsvendor problem.

This approach involves certain distributional assumptions and leads to better parametric estimation and optimization solution compared with traditional approach.

Data-driven Newsvendor problem with non-i.i.d. data. Our work is not the first to analyze data-driven solutions without assuming that data are generated in an i.i.d. manner. [5] examine the newsvendor problem with stationary autoregressive demand in a dynamic forecast-based method using mean squared error (MSE). [33] also consider the autoregressive demand and developed a robust optimization approach for demand forecasting. [31] investigate newsvendor problem with non-stationary demand. The authors propose a DPFNN-Based quantile auto-regression method that is applicable to stationary and non-stationary time series demand. Although the authors provide asymptotic analysis of their method, they did not provide any finite-sample performance guarantees. [114] propose a distributionally robust framework for quantile prediction (equivalent to the newsvendor problem) when the data are generated independently but not identically distributed. The authors provide both asymptotic and finite-sample performance guarantees. However, they only focus on the case where the demand is a linear function of the feature and a random noise term. [76] investigates the multi-period newsvendor problem with non-stationary demand as well as a random shock. The authors propose a moving window ordering policy that proved to have a regret with the smallest possible rate under certain conditions. Our problem setting is different from [76] in three fold: first, we consider the contextual information associated with the random demand; second, we investigate the single-period instead of the multi-period newsvendor problem; third, [76] consider a random shock besides random demand.

General frameworks for learning the data-to-decision mapping Numerous studies have examined methods to construct a mapping from the contextual information to decision, based on historical data. One of such approaches is the prescriptive learning, first proposed by [18]. Motivated by various machine learning models, including k-nearest neighbors and kernel regression methods, the prescriptive learning method first approximates the conditional distribution of contextual uncertainties and subsequently solves the corresponding stochastic optimization problem to yield the optimal decision based on a given contextual information. Another method focuses on directly learning the data-to-decision mapping by applying machine learning techniques. [110] apply the deep learning models to learn the newsvendor problem. [44] and [9] propose the smart Predict-then-Optimize (SPO) framework that leverages the optimization problem structure while forming the loss function. Thus a data-to-decision mapping can be constructed by learning using the SPO loss. The SPO framework can be applied to linear and piece-wise linear problems, which include the newsvendor problem. [39] adopt an end-to-end learning framework which directly use the objective function in the optimization stage as the loss function for learning the data-to-decision mapping. [117] propose a different end-to-end learning framework for a multi-period inventory management problem.

Among all previously-mentioned frameworks, only the prescriptive learning proposed by [18] and the SPO proposed by [44] have performance guarantees (the performance guarantees for SPO is derived in a later work [9]), both in a the form of generalization bounds. [9] consider i.i.d. samples while [9] also consider samples that are stationary but have intertemporal dependence. Our work distinguish from the result of [9] by considering non-i.i.d. samples and differ from the result of [18] by considering unbounded demand process and non-stationary demand process.

Machine learning methods with non-i.i.d. data. The issue of non-i.i.d. data has been investigated in the machine learning literature, mostly for the tasks of predicting a response variable with given features. [104] investigate the generalization ability of learning algorithms with the assumptions that samples are drawn from a β-mixing or ϕ-mixing stationary stochastic process and that both data and loss function are bounded. However, these assumptions may not represent the real requirements in contextual decision making, in which the tail behavior of a random quantity, e.g., demand, may significantly impact the decisions. Our study accommodates unbounded uncertainties and analyze how their tail behaviors affect the learning performance. The boundedness was relaxed in another study as well ([101]), in which the performance guarantees for learning methods are derived using a stationary, univariate autoregressive (AR) model. The authors only consider this specific AR process model, and the proofs rely on controlling the Gaussian complexity of the AR model. For comparison, our study only requires a light-tail property instead of the knowledge of the Gaussian complexity, and our method accommodates a more general class of time series models.

As performance guarantees, the studies of the generalization bounds for non-stationary data generating processes has also been reported in the machine learning literature. [87] provide the generalization bounds for time series predictions with a non-stationary mixing stochastic process, where the non-stationarity is characterized by the concept of discrepancy. The authors derive the generalization bound based on Rademacher complexity. Our results in Section 5.5 distinguish from those of [87] in two aspects: First, [87] obtain generalization results based on Rademacher complexity, whereas we do not require knowledge about the Rademacher complexity and provide a bound based on McDiamid inequalities. Second, [87] quantify the non-stationarity using discrepancy, which is defined as the supreme difference between the expected path-dependent generalization error at two periods. However, this measure of the non-stationarity discrepancy is not entirely natural in the settings for operations management and is computationally intractable. Real operations generate a natural property in that, even in presence of non-stationarity, the more recent historical data are more informative about future contextual uncertainties. Our analysis exploits this property to derive the generalization bounds when the non-stationarity is moderate, whereas we do not make structural assumptions on the specific forms of non-stationarity.

5.3 Learning the Contextual Newsvendor Problem

In this section, we describe the learning framework for the data-driven contextual newsvendor problem. We denote the demand of a product by $d \in \mathbb{R}$ and $x \in \mathbb{R}^p$ denotes the vector of contextual information or features that may be informative for the demand distribution. For both x and d, historical observations $\{x_i\}_{i=1}^n$ and $\{d_i\}_{i=1}^n$ are respectively available. The available data are generated sequentially from real operations, where the index i denotes data generated from a period, e.g., a day or an hour. Therefore, $\{(x_i, d_i)\}_{i=1}^n$ can be viewed as realizations from a random process $\{(X_i, D_i)\}$. We let u denote the order decision within the decision space $\mathcal{U}$. Subsequently, in the i-th time period, there is a inventory cost $c(d_i, u)$ incurred, including both back-ordering and holding costs. The inventory cost takes the form of

$$c(d, u) := b(d - u)^+ + h(u - d)^+, \tag{5.1}$$

where b and h are unit back-ordering and holding costs, respectively.

At the beginning of the i-th time period, the newsvendor has access to the context x_i and makes an order decision accordingly. After the decision making, the demand of the day, d_i, is revealed. The goal of a decision-maker is to exploit the observed dataset $\{(x_i, d_i)\}_{i=1}^n$ to learn a data-to-decision mapping denoted by $h(\cdot)$, which can be considered as a mapping from $\mathcal{X}$ to the contextual decision space $\mathcal{U}$. In this work, we consider a general case where $h(\cdot)$ is from a general hypothesis class $\mathcal{H}$. This is different from the setting in [10] where the authors consider a linear hypothesis class. Then, when a new context in the $(n+1)$-th time period X_{n+1} is revealed, a decision can be made by simply evaluating the mapping h.

The hypothesis $h(\cdot)$ is often constructed by applying certain machine learning techniques to data samples $S := \{(x_1, d_1), \ldots, (x_n, d_n)\}$. Thus, in the remaining part of this work, we let $h_S(\cdot)$ denote the hypothesis constructed based on samples S. Candidates of such learning algorithms include various machine learning models, such as kernel regression and random forest.

Therefore, the contextual newsvendor problem can be viewed as a machine learning problem that adopts the quantile loss, which is equivalently to the newsvendor cost up to scaling. The essential of the problem is to identify the data-to-decision mapping h_S that minimizes the expected newsvendor cost:

$$\min_{h_S:\, \mathcal{X} \to \mathcal{U}} E[c(D_{n+1}, h_S(X_{n+1}))|S], \tag{5.2}$$

where the expectation is taking over the joint distribution of (X_{n+1}, D_{n+1}), conditioning on samples S. Since $E[c(D_{n+1}, h_S(X_{n+1}))|S]$ is unknown, the objective function is often approximated by the in-sample average loss $\frac{1}{n}\sum_{i=1}^n c(d_i, h_S(x_i))$. Therefore a data-to-decision

mapping can be identified by solving

$$\min_{h_S:\mathcal{X}\to\mathcal{U}} \frac{1}{n}\sum_{i=1}^{n} c(d_i, h_S(x_i)). \tag{5.3}$$

After a h_S has been obtained by solving (5.3), it is critical for a decision-maker to evaluate the performance of h_S on out-of-sample unseen data. This is in relation to the objective of making a decision for a future case, i.e., the objective of (5.2). To quantify such a performance, we evaluate the gap between $R_{\text{emp}}(h_S)$ and $R(h_S)$, i.e. $|R(h_S) - R_{\text{emp}}(h_S)|$, where $R_{\text{emp}}(h_S) := \frac{1}{n}\sum_{i=1}^{n} c(d_i, h_S(x_i))$ and $R(h_S) = E[c(D_{n+1}, h_S(X_{n+1}))|S]$ denote the in-sample cost and out-of-sample cost, respectively. This gap is also referred to as the generalization error. Note that this generalization error is itself a random variable since it involves uncertainties in a future period. In this study, we provide bounds of the generalization error with any desired probability.

We would like to highlight that, in this work, we discard two unrealistic but extensively adopted assumptions. The first one is the i.i.d. assumption which assumes that $\{(x_i, d_i)\}_{i=1}^{n}$ are generated in an i.i.d. manner. However, in real contextual decision making systems, the i.i.d. assumption is rarely satisfied because the system often has a time-series nature. Dependence across pairs may exist. For example, demand is often predicted using previous sales records. Besides, as there is often a volatile environment, the demand process is often non-stationary. Therefore, in this study, we discard the unrealistic but extensively adopted i.i.d. assumption and consider inter-temporal dependence and moderate non-stationarity. The other assumption that is often adopted when investigating the out-of-sample performance of the demand process is that both demand and the contexts are bounded with a known upper-bound ([10, 101, 87]). However, it is difficult to impose such a known upper bound of uncertainties arising in real operations. Therefore, instead of boundedness, we use a weaker assumption that the demand adopts a light-tailed distribution. This assumption is more general and is satisfied by many common demand distributions, for example, Gaussian distribution.

5.4 Performance Guarantees under Intertemporal Dependence

In this section, we first relax the "independent" assumption and focus on intertemporal dependence when provide an out-of-sample performance guarantee. We further assume that the intertemporal dependence decays as the distance of two time periods gets further apart, which is quite natural. Intuitively, this assumption describes that the future depends more on data in the recent history than on data in the older history. One way to formally capture this property is through the concept of "mixing" for a stationary stochastic process ([40]). To fix our ideas, we state the definition for stationarity and mixing.

Definition 5.4.1 (Mixing)**.** Let $\{Z_t\}_{t=-\infty}^{\infty}$ be a stationary sequence of random variables. For any $i, j \in \mathbb{Z} \cap \{-\infty, +\infty\}$, let σ_i^j denote the σ-algebra generated by the random variables $Z_k, i < k < j$. Then, for any positive integer k, the ϕ-mixing coefficients of the stochastic process Z are defined as

$$\phi(k) = \sup_{\substack{n \\ A \in \sigma_{n+k}^{\infty} \\ B \in \sigma_{-\infty}^{n}}} \left| \Pr[A|B] - \Pr[A] \right| \tag{5.4}$$

and Z is said to be ϕ-mixing if $\phi(k) \to 0$ as $k \to \infty$.

Definition 5.4.2 (Stationarity)**.** A sequence of random variables $\{Z_t\}_{t=-\infty}^{\infty}$ is said to be stationary if for any t and non-negative integers m and k, the random vectors $(Z_t, \ldots, Z_t+m)$ and $(Z_{t+k}, \ldots, Z_{t+k+m})$ have the same distribution.

Definition 5.4.3 (Mixing)**.** Let $\{Z_t\}_{t=-\infty}^{\infty}$ be a stationary sequence of random variables. For any $i, j \in \mathbb{Z} \cap \{-\infty, +\infty\}$, let σ_i^j denote the σ-algebra generated by the random variables $Z_k, i < k < j$. Then, for any positive integer k, the ϕ-mixing coefficients of the stochastic process Z are defined as

$$\phi(k) = \sup_{\substack{n \\ A \in \sigma_{n+k}^{\infty} \\ B \in \sigma_{-\infty}^{n}}} \left| \Pr[A|B] - \Pr[A] \right| \tag{5.5}$$

and Z is said to be ϕ-mixing if $\phi(k) \to 0$ as $k \to \infty$.

Commonly used form of the mixing coefficients include algebraic mixing and exponentially mixing. The algebraically mixing means that there exist real numbers $\phi_0 > 0$ and $\gamma > 0$ such that $\phi(k) = \phi_0 k^{-\gamma}$ for all k while the exponentially mixing denotes the case where there exists $\phi_0 > 0$, s, and k such that $\phi(k) = \phi_0 s^k$. Intuitively speaking, if an contextual uncertainty process is ϕ-mixing, then its intertemporal dependence from two time periods, say t_1 and t_2 decays as the interval between t_1 and t_2 increases. In the remaining part of this section, we adopt the following assumption:

Assumption 8 (Stationarity and mixing property)**.** *The data sequence $\{(x_i, d_i)\}_{i=1}^n$ is a stationary and ϕ-mixing process $\{(X_t, D_t)\}_{t=-\infty}^{\infty}$.*

As an illustration, the case of i.i.d. data is a special case that satisfies Assumption 8, so as data generated by a bounded auto-regressive (AR) process and a Markov modulated process (MMP). For more commonly used examples that satisfy Assumption 8, we defer to Section C.1.

We also relax the typical assumption of boundedness of demand as it is rarely bounded in real-world applications, and even if they are bounded, it is unrealistic to identify the exact upper bound for future unseen demand. Therefore, instead of the boundedness assumption,

we extend to assume that $\{(X_t, D_t)\}_{t=-\infty}^{\infty}$ are light-tailed distributed with respect to the maximum norm $\|\cdot\|_\infty$.

Assumption 9. *(Light-tailed distribution) There exists a scalar $a > 1$, such that*

$$A := E[exp((\|(X_t, D_t)\|_\infty)^a)] < \infty. \tag{5.6}$$

This light-tailed assumption requires the tail of the distribution decay at an exponential rate. Under the light-tailed assumption, we provide the generalization bound for unbounded, mixing uncertainty processes.

5.4.1 Generalization Bound with Unbounded Demand Processes

We further assume the uniform stability of the machine learning algorithms adopted. In the following, we introduce the definition of uniform stability of a learning algorithm.

Definition 5.4.4. (Uniform Stability) An algorithm has uniform stability $\hat{\beta}$ with respect to the loss function c if the following holds

$$\forall S \in \mathcal{X}^n \times \mathcal{Y}^n, \forall i \in \{1, \ldots, n\}, \quad \sup_{x,y} |l(h_S, (x, y)) - l(h_{S^{\setminus i}}, (x, y))| \leq \hat{\beta}(n), \tag{5.7}$$

where n denotes the sample size, $S^{\setminus i}$ is defined as the sample S with the ith point removed, and $l(h_S, (x, y)) := c(y, h_S(x))$ is the loss function.

The stability coefficient $\hat{\beta}$ is a function of sample size n and an algorithm is said to be uniformly stable when $\hat{\beta}(n)$ decreases as $1/n$.

In the newsvendor problem, the cost function c is defined as 5.1. Thus, loss function is convex and σ-Lipschitz with $\sigma := \max\{b, h\}$. For example, if the hypothesis class is a reproducing kernel Hilbert space with kernel bounded by κ, then a L_2-regularized learning algorithm is uniformly stable with $\beta = \frac{\sigma^2\kappa^2}{2\lambda n}$ (Theorem 22 in [25]). Then Theorem 5.4.1 presents the generalization bound when data are generated by stationary mixing process on an unbounded support.

Theorem 5.4.1. *Suppose h_S is returned by a $\hat{\beta}(n)$-stable algorithm trained on data sample S and suppose that S is drawn from a stationary ϕ-mixing distribution satisfying Assumption 9. Then for any $m \in \{0, \cdots, n\}$ and any $\varepsilon > 0$, the following generalization bound holds:*

$$\begin{aligned}
&\Pr_S\left[\left|R(h_S) - R_{emp}(h_S)\right| > \varepsilon + (6m+2)\hat{\beta}(n) + 6M(n)(b \vee h)\phi(m)\right] \\
&\leq 2\exp\Big(\frac{-2\varepsilon^2(1 + 2\sum_{i=1}^{n}\phi(i))^{-2}}{n(2(m+2)\hat{\beta}(n) + 2M(n)(b \vee h)\phi(m) + \frac{M(n)(b\vee h)}{n})^2}\Big) + A\exp(\ln n - M^a(n))
\end{aligned} \tag{5.8}$$

where $\hat{\beta}(n)$ is the uniformly stable coefficient, $M(n)$ is a large number as a function of sample size that can be tuned according to different mixing coefficients.

The generalization bound stated in Theorem 5.4.1 is derived based on a coupling sample which is defined as the original sample truncated by a sample-dependent large number $M(n)$. By applying existing generalization result to the coupled truncated copy and carefully adjusting the truncated threshold, we can bound the generalization error from an unbounded mixing process.

Before introducing the proof of Theorem 5.4.1, we first construct the coupling example. In the remaining part of this paper, we let Z_t denote the abbreviation of the contextual process (X_t, D_t). Similarly, we let z_i denote the abbreviation of the observations (x_i, d_i). Then the training sample can be represented as $S = \{z_1, \ldots, z_n\}$. For any training sample S, we define a coupling sample

$$\tilde{S} := \{\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_i, \ldots, \tilde{z}_n\} \tag{5.9}$$

with respect to the sample

$$S := \{z_1, z_2, \ldots, z_i, \ldots, z_n\} \tag{5.10}$$

where $\tilde{z}_i$ is defined as

$$\tilde{z}_i = \begin{cases} \tilde{z}_i = z_i, & \text{if } \|z_i\|_\infty \leq M(n) \\ (\tilde{z}_i)_{k,l} = \min\{(z_i)_{k,l}, M(n)\} & \text{o.w.} \end{cases} \tag{5.11}$$

The coupling sample $\tilde{S}$ is defined by truncating the original sample S with respect to $M(n)$ which is carefully calibrated and is an increasing function of sample size n. It is easy to check that, if $\{z_i\}$ is generated from a ϕ-mixing process, then $\{\tilde{z}_i\}$ is also ϕ-mixing, due to the fact that $\sigma_{n+k}^{\infty}(\tilde{Z}) \subset \sigma_{n+k}^{\infty}(Z)$. Then the following lemma bounds the probability of S being different of $\tilde{S}$.

Lemma 5.4.2. *Suppose S is a n-sized sample set generated by ϕ-mixing process, and the marginal distribution is light-tailed as described in Assumption 9. Then the following holds:*

$$\Pr_S(\tilde{S} \neq S) \leq \frac{nA}{\exp(M^a(n))} \tag{5.12}$$

The proof of Lemma 5.4.2 can be found in the Appendix. Based on the result stated in Lemma 5.4.2, we deliver the proof of Theorem 5.4.1.

Proof of Theorem 5.4.1. Consider the truncated coupling we constructed earlier in this section, the loss function $l(h_S, z_i)$ is upper bounded by $M(n)(b \vee h)$ due to its Lipschitz property.

By applying the generalization bound from [104] for mixing process with bounded loss function, we have

$$\begin{aligned}&\Pr_S\left[\left|R(h_S)-R_{emp}(h_s)\right|>\varepsilon+(6m+2)\hat{\beta}(n)+6M(n)(b\vee h)\phi(b)\right]\\&\leq 2\exp\Big(\frac{-2\varepsilon^2(1+2\sum_{i=1}^{n}\phi(i))^{-2}}{n((m+2)2\hat{\beta}(n)+2M(n)(b\vee h)\phi(m)+\frac{M(n)(b\vee h)}{n})^2}\Big)\end{aligned}$$

for any $m\in\{0,\cdots,n\}$ and any $\varepsilon>0$.

Then the desired result can be achieved by applying Lemma 5.4.2 to bound the probability of $S\neq\tilde{S}$. Thus, $\Pr_S\left[|R(h_S)-R_{\text{emp}}(h_S)|\right]\leq\Pr_{\tilde{S}}\left[|R(h_{\tilde{S}})-R_{\text{emp}}(h_{\tilde{S}})|\right]+\Pr(S\neq\tilde{S})$ leads to the result of Theorem 5.4.1. □

With more knowledge of mixing coefficient and uniformly stable coefficient, the generalization bound can be further tightened by carefully choosing m and $M(n)$. In particular, under the case of exponential mixing data (when $\phi(k)=\phi_0 s^k$) and uniformly stable algorithm (with $\hat{\beta}(n)=\frac{\hat{\beta}}{n}$), an optimized generalization bound is

Theorem 5.4.3. *Suppose the light-tail parameter $a=1$ and h_S is returned by a $\frac{\beta_0}{n}$-stable algorithm trained on a sample S drawn from an exponentially ϕ-mixing stationary distribution,i.e. $\phi(k)=\phi_0 s^k$ for some $0<s<1$. The marginal distribution of sample set satisfies the light-tail assumption, i.e. Assumption 9 as well. Hence, for any $\varepsilon>0$, and n large enough, the following generalization bound holds by setting $M(n)=n^{\frac{1}{3}}$:*

$$\begin{aligned}&\Pr_S\left[\left|R(h_S)-R_{emp}(h_S)\right|>\varepsilon+(6m^*+2)\beta_0/n+6\beta_0/(n\ln\frac{1}{s})\right]\\&\leq 2\exp\Big(-\frac{2\varepsilon^2(1-s)^2}{(1-s+2\phi_0)^2}\frac{n}{(2\beta_0(m^*+2)+2\frac{\beta_0}{\ln\frac{1}{s}}+(b\vee h)n^{1/3})^2}\Big)+A\exp\big(\ln n-n^{1/3}\big).\end{aligned}\tag{5.13}$$

where $m^*=\frac{\beta_0}{b_0\phi_0 n^{1+1/3}\ln\frac{1}{s}}$.

Asymptotically, the right-hand side converges with rate $\sim\exp(-n^{a/(a+2)})$, while the error term on the left hand side converges by rate $\sim\frac{1}{n}$.

5.4.2 Comparison with Results from [10]

In this section, we adopt the same assumptions as [10]. We summarize and list these assumptions as the following:

Assumption 10. *Key assumptions adopted in [10]:*

1. *The samples* $\{(x_i, d_i)\}_{i=1}^n$ *are i.i.d. in* i.
2. *The contextual information,* X*, is bounded by* X_{max}*, i.e.,* $X \leq X_{max}$.
3. *Demand,* D*, is bounded by* $\bar{D}$*, i.e.,* $D \leq \bar{D}$.
4. *Demand,* D*, linearly depends on* x*, i.e.,* $D = h^T x + \varepsilon$*, where* ε *is a random noise.*

We consider the ERM-2 learning algorithm proposed in [10] for further comparison. The ERM-2 learning algorithm can be viewed as the quantile regression with $L2$-regularization. Specifically, a data-to-decision mapping $h_S(x)$ can be achieved by solving

$$\min_{h_S \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} c(d_i, h_S(x_i)) + \lambda \|h_S\|, \tag{5.14}$$

where $\mathcal{H}$ denotes the set of linear functions and each $h_S \in \mathcal{H}$ corresponding to a linear function of x, say, $h^T x$ for some $h \in \mathbb{R}^p$. λ represents a hyper parameter of regularization and $\|h_S\| := \|h\|_2$ denotes the $L2$-norm of the linear function coefficient vector h. As a quantile regression with L_2 regularization, the ERM-2 algorithm is proved to be uniformly stable with coefficient $\hat{\beta} \leq \frac{\sigma^2 \kappa^2}{2\lambda n}$, where $\sigma = b \vee h$, $\kappa = X_{\max}\sqrt{p}$ and λ denotes the coefficient for regularization ([25]).

Corollary 5.4.3.1 (Learning Newsvendor problem with i.i.d. data)**.** *Suppose Assumption 10 holds, based on Theorem 12 of [25], Theorem 6 in [10] can be further recovered as:*

$$\Pr_S(|R(h_S) - R_{emp}(h_S)| > \varepsilon + \frac{(b \vee h)^2 X_{max}^2 p}{n\lambda}) \leq 2 \exp\big(-\frac{2n\varepsilon^2}{(\frac{2(b\vee h)^2 X_{max}^2 p}{n\lambda} + \bar{D}(b \vee h)))^2}\big), \tag{5.15}$$

where n *denotes the sample size.*

However, the i.i.d. assumption on $\{(x_1, d_1), \ldots, (x_n, d_n)\}$ may limit the scope of applications in real operations. For example, an important feature that can be used to inform the future demand uncertainty is the historical demand. That is, to predict the stochastic demand d_j in the j-th time period, it is often helpful to include $d_1, \ldots, d_{j-1}$, demands observed from the previous time period as feature, which means that the contextual information x_j may contain d_{j-1}. In the simplest case is the AR(1) model, where data samples $\{(x_i, d_i)\}_{i=1}^n$ take the form of $\{(d_{i-1}, d_i)\}_{i=1}^n$. It is evident that, although used as a motivating example in [10], samples $\{(x_i, d_i))\}_{i=1}^n$ that generated by the AR(1) model do not satisfy the i.i.d. assumption. Therefore, the performance guarantees provided by (5.15) are no longer valid. Note that the AR(1) model satisfies the ϕ-mixing property defined earlier. In fact, for the AR(1) model, we can find ϕ_0 and γ such that $\phi(k) \leq \phi_0 k^{-\gamma}$.

Therefore, we relaxed the first term of Assumption 10 and instead consider the case in which the demand-feature process is stationary and algebraically ϕ-mixing, i.e., $\phi(k) =$

$\phi_0 k^{-\gamma}$. Together with the remaining term (2.–4.) of Assumption 10, Corollary 5.4.3.2 provides an alternative performance guarantee.

Corollary 5.4.3.2. *(Learning newsvendor problem with algebraically ϕ-mixing data) When learning the newsvendor problem, suppose $2.-4.$ of Assumption 10 hold and data samples are generated by an algebraically ϕ-mixing process, i.e., $\phi(k) = \phi_0 k^{-\gamma}$, then the following performance guarantee holds:*

$$\begin{aligned}\Pr_S\big(\big|R(h_S) - R_{emp}(h_S)\big| > \varepsilon + (6u^*+2)\frac{(b\vee h)^2 X_{max}^2 p}{2\lambda n} + 6(\gamma+1)\bar{D}(b\vee h)\phi(u^*)\big)\\ \leq 2\exp\big(-\frac{2n\varepsilon^2(1+2\phi_0\gamma/(\gamma-1))^{-2}}{\big(\frac{(u^*+2)(b\vee h)^2X_{max}^2 p}{\lambda} + 2n(\gamma+1)\bar{D}(b\vee h)\phi(u^*) + \bar{D}(b\vee h)\big)^2}\big), \quad (5.16)\end{aligned}$$

where $u^* = \big(\frac{(b\vee h)X_{max}^2 p}{2\lambda n\gamma\phi_0\bar{D}}\big)^{-1/(\gamma+1)}$.

Compared to (5.15), (5.16) is looser, which is quite intuitive, since correlation among samples leads to an inaccurate learning. In addition, if the demand-process is i.i.d., then $u^* = 0$, $\phi(k) = 0$ for all k and $\phi_0 = 0$ and (5.16) recovers (5.15).

5.5 Performance Guarantees under Moderate Non-stationarity

In addition to relax the "independent" assumption, we further relax the assumption of stationarity, which is an adaption of "identically distributed". To simplify the analysis, we instead assume that the demand process adopts a Markovian structure. To take non-stationarities into consideration, the corresponding Markov chain can be nonhomogeneous, which means the transition kernel may vary across different time period. We formally state these assumptions as follows:

Assumption 11. *Suppose $\{z_i\}_{i=0}^n$ is generated by a (nonhomogeneous) Markov chain $\{Z_t\}_{t=0}^n$. Let $p_0(\cdot)$ denote the initial probability distribution of Z_0 and $p_i(\cdot|\cdot)$, $1 \leq i \leq n-1$ denote its transition kernel. Then*

$$\mathbb{P}\{(Z_1,\ldots,Z_i) = (z_1,\ldots,z_i)\} = p_0(z_1)\Pi_{j=1}^{i-1}p_j(z_{j+1}|z_j). \quad (5.17)$$

Suppose the contraction property holds and the i-th contraction coefficient is denoted by

$$\theta_i = \sup_{z',z''\in\mathcal{S}} \|p_i(\cdot|z') - p_i(\cdot|z'')\|_{TV}. \quad (5.18)$$

We consider the measure space $(\Pi_{k=1}^n\Omega_k, \Pi_{k=1}^n\mathcal{F})$ on $\{Z_1\ldots,Z_n\}_{i=1}^n$, and denote P as a measure on this space with P_j as its marginal on $(\Pi_{k=1}^j\Omega_k, \Pi_{k=1}^j\mathcal{F})$. Then, for any $i < j$, the following lemma holds:

Lemma 5.5.1. *Suppose Assumption 11 holds, then we have*

$$\|P_j(\cdot|\mathcal{F}_i) - P_j\|_{TV} \leq \Pi_{l=i+1}^{j}\theta_l \tag{5.19}$$

where θ_l is the contraction coefficient defined in (5.18).

Proof of Lemma 5.5.1. Considering the definition of θ_i and the quantity η_i^j which is defined as

$$(\eta_i^j)^2 = \sup_{z',z''\in S} \|P_j(Z_j|Z_i = z') - P_j(Z_i|Z_i = z'')\|_{TV},$$

by the result (Equation 2.9) of [128], we have $(\eta_i^j)^2 \leq \Pi_{l=i+1}^{j}\theta_l$, which leads to the desired results. □

Similar to previous parts, we let $S = \{(x_i, d_i)\}_{i=1}^n$ denote the training samples and consider a $\hat{\beta}$-stable learning algorithm that output a hypothesis hypothesis h_S. The empirical loss is defined as $R_{emp}(h_S) = \sum_{i=1}^n c(d_i, h_S(x_i))$. For the out-of-sample cost, instead of the expected cost for (X_{n+1}, D_{n+1}), i.e., $R(h_S) = \mathbb{E}[c(D_{n+1}, h_S(X_{n+1}))]$, we consider a more general case in which the out-of-sample cost might be evaluated at period $n + r + 1$, i.e.,

$$R_{n+r+1}(h_S) = \mathbb{E}[c(D_{n+r+1}, h_S(X_{n+r+1}))]. \tag{5.20}$$

Note that $R(h_S)$ defined in Section 5.3 is a special case of (5.20) with $r = 0$. In the remaining part of this section, we define $\Phi(S) = R_{emp}(h_S) - R_{n+r+1}(h_S)$ and consider the case with bounded loss function for simplicity. Then Theorem 5.5.2 demonstrates the generalization bound:

Theorem 5.5.2. *Suppose the cost function c is upper-bounded by M and we define $M_n := \max_{1\leq i\leq n-1}(1 + \theta_i + \theta_i\theta_{i+1} + \cdots + \theta_i \ldots \theta_{n-1})$, then $\forall b \in \{0, 1, \ldots, \lfloor \frac{n}{2} \rfloor\}$*

$$\mathbb{P}\{|R_{emp}(h_S) - R_{n+r+1}(h_S)| \geq \varepsilon + 4b\hat{\beta} + \frac{2M}{n-2b}\sum_{i=b+1}^{n-b}(\Pi_{l=i-b+1}^{i}\theta_l + \Pi_{l=i+1}^{i+b}\theta_l + \Pi_{l=n+1}^{n+r+1}\theta_l) + \Delta\}$$

$$\leq 2\exp\Big(\frac{-\varepsilon^2 M_n^{-2}}{2n((4+4b)\hat{\beta} + 2\frac{M}{n} + 2M\Pi_{l=n-b+1}^{n+r+1}\theta_l)^2}\Big) \tag{5.21}$$

In order to prove Theorem 5.5.2, we first show the Lipschitz property of Φ with respect to Hamming metric in the following lemma:

Lemma 5.5.3. *Suppose $S = \{x_1, \ldots, x_i, \ldots, x_n\}$ and $S^i = \{x_1, \ldots, x_i', \ldots, x_n\}$ are two sequence generated by a contractive Markov chain with only on different sample z_i and z_i'. Then $\Phi(S)$ is Lipschitz with respect to the hamming metric:*

$$|\Phi(S) - \Phi(S^i)| \leq 4\hat{\beta} + 2\frac{M}{n} + 2M\Pi_{l=n+1}^{n+r+1}\theta_i \tag{5.22}$$

Since the Markov chain is non-stationary, we use the total variation of the marginal distributions for two different time steps, denoted by $\delta_{i,j} := \|P_i - P_j\|_{TV}$, to quantify the nonstationarity. For stationary process as a special case, $\delta_{i,j}$ is zero, so that there is no nonstationarity. We further adopt the following assumption to make sure that the nonstationarity does not explode.

Assumption 12. *Suppose* $\Delta_m^{m+k} = \frac{1}{m}\sum_{i=1}^{m} \delta_{i,m+k}$ *is bounded by* Δ *for any* m, k *and* $\Delta_m^{m+k} \to 0$ *as* $m \to \infty$.

With this characterization of nonstationarity, we then provide an upper-bound for $\mathbb{E}[\Phi(S)]$ with the following lemma.

Lemma 5.5.4. *With Assumption 11 and Assumption 12,* $\mathbb{E}[\Phi(S)]$ *can be bounded by the following results*

$$\mathbb{E}[\Phi(S)] \leq 4b\hat{\beta} + 2M(\Pi_{l=i-b+1}^{i}\theta_l + \Pi_{l=i+1}^{i+b}\theta_l + \Pi_{l=m+1}^{m+r+1}\theta_l) + \Delta \tag{5.23}$$

The proof of Lemma 5.5.3 and Lemma 5.5.4 can be found in the Appendix.

Proof of Theorem 5.5.2. The result of Theorem 5.5.2 can be achieved by Theorem 1.2 in [84] together with Lemma 5.5.3 and Lemma 5.5.4. □

Moreover, if the transition kernel $p_i(\cdot|\cdot)$ varies periodically, i.e. $p_i(\cdot|\cdot) = p_{i \bmod T}(\cdot|\cdot), \forall i$, the result stated in Theorem 5.5.2 can be further simplified by plugging in $M_n := \max_{1\leq i\leq T}(1+\theta_i + \theta_i\theta_{i+1} + \cdots + \theta_i \ldots \theta_{n-1})$ and $\Pi_{l=i}^{j}\theta_l \leq \Theta_{i,j} := \max_{1\leq t\leq T} \theta_t\theta_{t+1}\ldots\theta_{t+j-i}$.

Corollary 5.5.4.1. *Suppose the items 2-4 of Assumption 10 hold and samples* $\{(x_i, d_i)\}_{i=1}^{n}$ *are generated from a nonhomogeneous Markov chain, the performance guarantee of the* $n+1$ *period can be stated as*

$$\begin{aligned}\mathbb{P}\{|R_{emp}(h_S) - R_{n+1}(h_S)| \geq \varepsilon + 2m\frac{(b\vee h)^2 X_{max}^2 p}{n\lambda} + \frac{2b\vee hD_{max}}{n-2m}\sum_{i=b+1}^{n-m}(\Theta_{i-m+1,n-m} + \Theta_{i+1}^{i+m} + \theta_{n+1}) + \Delta\} \\ \leq 2\exp\Big(\frac{-\varepsilon^2 M_n^{-2}}{2n((2+2m)\frac{(b\vee h)^2 X_{max}^2 p}{n\lambda} + 2\frac{b\vee hD_{max}}{n} + 2b\vee hD_{max}\Theta_{n-b+2,n+1})^2}\Big).\end{aligned} \tag{5.24}$$

Noted that the result stated in (5.24) depends on the contraction coefficients θ_t for each time period t. Generally speaking, $\Delta \sim O(1)$. Therefore, (5.24) is generally looser compared to (5.16) with an inevitable error term Δ in the left-hand side owing to the nonhomogeneity of the Markov chain. This result is also quite intuitive, since in a non-stationary environment, it is more difficult to learn a good data-to-decision mapping from historical data.

5.6 Numerical Experiments

In this section, we present the numerical benefits of including intertemporal dependent features when learning data-to-decision mappings. More specifically, we consider the features that are generated from the historical observations of the contextual uncertainty, i.e., in time period i, the feature vector x_i contains or depends on previous observations of demand $d_1 \ldots, d_{i-1}$. We show that including such features in learning contextual decisions can significantly improve the decision performance via machine learning tools, thereby confirming our premise to build rigorous performance guarantees for machine learning tools using intertemporally dependent data.

We consider the newsvendor example stated in Section C.1 with real-world sales data from a large e-retailer in China. Given daily demand data and external information records, the objective is to determine the optimal stock level according to the newsvendor problem. Our data include the sales and external information of a duration of 917 days, from January 1, 2016 to July, 6, 2018. The dataset consists of the data of seven stock keeping units (SKU) among eight distribution centers (DC). There are 51352 total samples for both daily sales and external information. Among all the available data, the training and test sets are split according to time periods, where the data from the most recent 217 days are used as the test period, and the data from the previous 700 days constitute the training set. Therefore, there are 44800 samples in the training set and 6552 samples in the test set.

We consider two different learning algorithms, the linear model and gradient boosting, as representatives of parametric and non-parametric methods, respectively, for learning the data-to-decision mapping. Specifically, the data-to-decision mapping, h_S, is determined by a linear function or a gradient boosting model, respectively.

We consider two different sets of features. The first set of features only includes external information, which are not temporally correlated. External information includes SKU- and DC-level information, such as the dummy variables generated based on SKU ID and DC ID. The second set of features includes the features that are generated from historical observations, such as the so-called auto-regressive (AR) and moving average (MA) features. The AR features denote historical sales of the past m days, for some value of m. In this experiment, we take $m = 7$. The MA features denote the average sales over a time period in the past. Particularly, for each time period i, the MA features include 5 average demands over a 30-day period counting back from the period $i - 7$. Specifically, the MA features consist of the average demand from period $i - 37$ to $i - 8$, from $i - 67$ to $i - 38$, from $i - 97$ to $i - 68$, from $i - 127$ to $i - 98$, and from $i - 157$ to $i - 128$. In the remaining part of this section, we denote the two sets of features as EI and HO features, respectively, for abbreviation. Notably, the EI features are often considered as generated independently and identically distributed. However, the HO features are generated in time-series manner and are intertemporally dependent.

In the numerical experiments, we consider the newsvendor problem in which the parameter b, the cost of unit back-order cost, takes the value of $b = 9$ and the parameter, h, the cost of unit holding cost, takes the value of $h = 1$. The test set performances are evaluated as the averaged test set inventory cost, which includes both holding and back-order costs.

In Table 5.1 demonstrate test set performances of the LR models with different sets of features. For each of the methods, we summarize the features included, the average newsvendor cost over the test set, and the relative saving compared to the model that only involves EI features. Not surprisingly, the best linear model is the one that includes both EI and HO features. The best linear model achieves a substantial improvement (32.72%) compared to the model that only involves external information. The results also indicate that a similar performance (30.56% improvement) can be achieved with HO features only.

Method	External information	Historical Observations	Average Cost	Relative Saving
LR-EI	Yes	No	6.02	N/A
LR-HO	No	Yes	4.18	30.56%
LR-ALL	Yes	Yes	4.05	32.72%

Table 5.1: Summary of Results Using Linear Methods

Method	External information	Historical Observations	Average Cost	Relative Saving
GB-EI	Yes	No	4.68	N/A
GB-HO	No	Yes	3.96	15.38%
GB-ALL	Yes	Yes	3.90	16.67%

Table 5.2: Summary of Results Using Gradient Boosting

Table 5.2 summarizes the results obtained when the data-to-decision mapping is learned using gradient boosting. The best out-of-sample performance is also achieved by the model including both EI and HO features. The gradient boosting model that only involves HO features also achieves a substantial improvement (15.38%), whereas the best model achieves a similar improvement of 32.72%.

To summarize, the results in Tables 5.1 and Table 5.1 demonstrate that, there are substantial benefits of considering intertemporally dependent features into consideration, for the two learning algorithms, which can be viewed as representatives of the parametric and nonparametric, and the linear and nonlinear methods, respectively. Owing to the significant benefits of including intertemporally dependent features, it is crucial to develop valid performance guarantees accordingly.

5.7 Conclusion

This study focusing on investigating the performance guarantees of learning a data-to-decision mapping for the newsvendor problem. The performances of such a framework on unseen future scenarios is often studied relying on assumptions that are highly restrictive in real contextual environments. Such assumptions include that the observed data are i.i.d. and adopt values in a bounded range.

By relaxing such unrealistic assumptions, we provide theoretically justified performance guarantees for two more realistic settings. First, we relax the bounded assumption and develop generalization bounds for a case in which the data samples are generated from a stationary but intertemporally dependent process characterized by the ϕ-mixing property. The ϕ-mixing property is validated using models that are commonly used to characterize contextual uncertainties and we provide a general recipe to evaluate this property in practice. Then we further relax the stationary assumption and develop generalization bounds for scenarios in which the underlying data samples are generated through a nonhomogeneous Markov process.

Numerical experiments are conducted to demonstrate the benefit of including intertemporally dependent contextual information when learning data-to-decision mapping. The numerical results further validate the significance of studying a learning framework considering intertemporal correlations.

Bibliography

[1] Soroosh Shafieezadeh Abadeh, Peyman Mohajerin Mohajerin Esfahani, and Daniel Kuhn. "Distributionally robust logistic regression". In: *Advances in Neural Information Processing Systems*. 2015, pp. 1576–1584.

[2] Akshay Agrawal et al. "Differentiable Convex Optimization Layers". In: *Advances in Neural Information Processing Systems*. 2019, pp. 9558–9570.

[3] Hesam Ahmadi and Uday V Shanbhag. "Data-driven first-order methods for misspecified convex optimization problems: Global convergence and rate estimates". In: *53rd IEEE Conference on Decision and Control*. IEEE. 2014, pp. 4228–4233.

[4] Ravindra K Ahuja, Thomas L Magnanti, and James B Orlin. "Network flows". In: (1988).

[5] Layth C Alwan et al. "The dynamic newsvendor model with correlated demand". In: *Decision Sciences* 47.1 (2016), pp. 11–30.

[6] Luca Ambrogioni et al. "The kernel mixture network: A nonparametric method for conditional density estimation of continuous random variables". In: *arXiv preprint arXiv:1705.07111* (2017).

[7] Brandon Amos and J Zico Kolter. "Optnet: Differentiable optimization as a layer in neural networks". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 136–145.

[8] Krishna B Athreya and Sastry G Pantula. "Mixing properties of Harris chains and autoregressive processes". In: *Journal of applied probability* 23.4 (1986), pp. 880–892.

[9] Othman El Balghiti et al. "Generalization Bounds in the Predict-then-Optimize Framework". In: *arXiv preprint arXiv:1905.11488* (2019).

[10] Gah-Yi Ban and Cynthia Rudin. "The big data newsvendor: Practical insights from machine learning". In: *Operations Research* 67.1 (2019), pp. 90–108.

[11] Peter L Bartlett and Shahar Mendelson. "Rademacher and Gaussian complexities: Risk bounds and structural results". In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 463–482.

[12] Basel Committee on Banking Supervision. *Revisions to the Basel II market risk framework*. Bank for International Settlements, 2009.

[13] Güzin Bayraksan and David K Love. "Data-driven stochastic programming using phi-divergences". In: *The Operations Research Revolution*. INFORMS, 2015, pp. 1–19.

[14] Yoshua Bengio. "Using a financial training criterion rather than a prediction criterion". In: *International Journal of Neural Systems* 8.04 (1997), pp. 433–443.

[15] Quentin Berthet et al. "Learning with differentiable perturbed optimizers". In: *arXiv preprint arXiv:2002.08676* (2020).

[16] Dimitris Bertsimas, Jack Dunn, and Nishanth Mundru. "Optimal prescriptive trees". In: *INFORMS Journal on Optimization* 1.2 (2019), pp. 164–183.

[17] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. "Robust sample average approximation". In: *Mathematical Programming* (2017), pp. 1–66.

[18] Dimitris Bertsimas and Nathan Kallus. "From predictive to prescriptive analytics". In: *Management Science* 66.3 (2020), pp. 1025–1044.

[19] Dimitris Bertsimas and Christopher McCord. "From predictions to prescriptions in multistage optimization problems". In: *arXiv preprint arXiv:1904.11637* (2019).

[20] Anna-Lena Beutel and Stefan Minner. "Safety stock planning under causal demand forecasting". In: *International Journal of Production Economics* 140.2 (2012), pp. 637–645.

[21] Jose Blanchet, Yang Kang, and Karthyek Murthy. "Robust wasserstein profile inference and applications to machine learning". In: *Journal of Applied Probability* 56.3 (2019), pp. 830–857.

[22] Jose Blanchet and Karthyek Murthy. "Quantifying distributional model risk via optimal transport". In: *Mathematics of Operations Research* 44.2 (2019), pp. 565–600.

[23] François Bolley, Arnaud Guillin, and Cédric Villani. "Quantitative concentration inequalities for empirical measures on non-compact spaces". In: *Probability Theory and Related Fields* 137.3-4 (2007), pp. 541–593.

[24] Joos-Hendrik Böse et al. "Probabilistic demand forecasting at scale". In: *Proceedings of the VLDB Endowment* 10.12 (2017), pp. 1694–1705.

[25] Olivier Bousquet and André Elisseeff. "Stability and generalization". In: *Journal of machine learning research* 2.Mar (2002), pp. 499–526.

[26] Andrea Braides. "A handbook of Gamma-convergence". In: *Handbook of Differential Equations: stationary partial differential equations*. Vol. 3. Elsevier, 2006, pp. 101–213.

[27] Andrea Braides et al. *Gamma-convergence for Beginners*. Vol. 22. Clarendon Press, 2002.

[28] Andrew Butler and Roy Kwon. "Integrating prediction in mean-variance portfolio optimization". In: *Available at SSRN 3788875* (2021).

[29] A Colin Cameron and Pravin K Trivedi. *Microeconometrics: methods and applications*. Cambridge university press, 2005.

[30] Ying Cao. *Data-driven Approaches to Inventory Management*. University of California, Berkeley, 2019.

[31] Ying Cao and Zuo-Jun Max Shen. "Quantile forecasting and data-driven inventory management under nonstationary demand". In: *Operations Research Letters* 47.6 (2019), pp. 465–472.

[32] Carla Cardinali. "Observation influence diagnostic of a data assimilation system". In: *Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. II)*. Springer, 2013, pp. 89–110.

[33] Emilio Carrizosa, Alba V Olivares-Nadal, and Pepa Ramırez-Cobo. "Robust newsvendor problem with autoregressive demand". In: *Computers & Operations Research* 68 (2016), pp. 123–133.

[34] Katherine Elizabeth Castellano and Andrew Dean Ho. "Contrasting OLS and quantile regression approaches to student "growth" percentiles". In: *Journal of Educational and Behavioral Statistics* 38.2 (2013), pp. 190–215.

[35] Anna Choromanska et al. "The loss surfaces of multilayer networks". In: *Artificial Intelligence and Statistics*. 2015, pp. 192–204.

[36] Leon Yang Chu, J George Shanthikumar, and Zuo-Jun Max Shen. "Solving operational statistics via a Bayesian analysis". In: *Operations Research Letters* 36.1 (2008), pp. 110–116.

[37] Yurii Aleksandrovich Davydov. "Mixing conditions for Markov chains". In: *Teoriya Veroyatnostei i ee Primeneniya* 18.2 (1973), pp. 321–338.

[38] Etienne De Klerk, Dick Den Hertog, and G Elabwabi. "On the complexity of optimization over the standard simplex". In: *European journal of operational research* 191.3 (2008), pp. 773–785.

[39] Priya Donti, Brandon Amos, and J Zico Kolter. "Task-based end-to-end model learning in stochastic optimization". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5484–5494.

[40] Paul Doukhan. *Mixing: properties and examples*. Vol. 85. Springer Science & Business Media, 2012.

[41] Richard Ehrhardt. "(s, S) policies for a dynamic inventory model with stochastic lead times". In: *Operations Research* 32.1 (1984), pp. 121–132.

[42] Kenneth J Ellis, Steven A Abrams, and William W Wong. "Monitoring childhood obesity: assessment of the weight/height2 index". In: *American journal of epidemiology* 150.9 (1999), pp. 939–946.

[43] Adam Elmachtoub, Jason Cheuk Nam Liang, and Ryan McNellis. "Decision trees for decision-making under the predict-then-optimize framework". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2858–2867.

[44] Adam N Elmachtoub and Paul Grigas. "Smart "predict, then optimize"". In: *Management Science* (2021).

[45] E Erdoğan and Garud Iyengar. "Ambiguous chance constrained problems and robust optimization". In: *Mathematical Programming* 107.1-2 (2006), pp. 37–61.

[46] Peyman Mohajerin Esfahani and Daniel Kuhn. "Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations". In: *Mathematical Programming* 171.1-2 (2018), pp. 115–166.

[47] Chenyou Fan et al. "Multi-horizon time series forecasting with temporal attention learning". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM. 2019, pp. 2527–2535.

[48] Hadi Fanaee-T and Joao Gama. "Event labeling combining ensemble detectors and background knowledge". In: *Progress in Artificial Intelligence* 2.2-3 (2014), pp. 113–127.

[49] Aaron Ferber et al. "Mipaal: Mixed integer program as a layer". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 02. 2020, pp. 1504–1511.

[50] Mark E Fetler. "A method for the construction of differentiated school norms". In: *Applied Measurement in Education* 4.1 (1991), pp. 53–66.

[51] Nicolas Fournier and Arnaud Guillin. "On the rate of convergence in Wasserstein distance of the empirical measure". In: *Probability Theory and Related Fields* 162.3-4 (2015), pp. 707–738.

[52] LLdiko E Frank and Jerome H Friedman. "A statistical view of some chemometrics regression tools". In: *Technometrics* 35.2 (1993), pp. 109–135.

[53] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, NY, USA, 2001.

[54] Guillermo Gallego and Özalp Özer. "Integrating replenishment decisions with advance demand information". In: *Management science* 47.10 (2001), pp. 1344–1360.

[55] Bolin Gao and Lacra Pavel. "On the properties of the softmax function with application in game theory and reinforcement learning". In: *arXiv preprint arXiv:1704.00805* (2017).

[56] Rui Gao, Xi Chen, and Anton J Kleywegt. "Wasserstein distributional robustness and regularization in statistical learning". In: *arXiv preprint arXiv:1712.06050* (2017).

[57] Rui Gao and Anton J Kleywegt. "Distributionally robust stochastic optimization with Wasserstein distance". In: *arXiv preprint arXiv:1604.02199* (2016).

[58] Paul Grigas, Meng Qi, et al. "Integrated Conditional Estimation-Optimization". In: *arXiv preprint arXiv:2110.12351* (2021).

[59] Vishal Gupta and Nathan Kallus. "Data pooling in stochastic optimization". In: *Management Science* (2021).

[60] Lauren Hannah, Warren Powell, and David M Blei. "Nonparametric density estimation for stochastic optimization with an observable state variable". In: *Advances in Neural Information Processing Systems*. 2010, pp. 820–828.

[61] Chin Pang Ho and Grani A Hanasusanto. *On Data-Driven Prescriptive Analytics with Side Information: A Regularized Nadaraya-Watson Approach*. Tech. rep. Technical report, March, 2019.

[62] Weihua Hu et al. "Does distributionally robust supervised learning give robust classifiers?" In: *arXiv preprint arXiv:1611.02041* (2016).

[63] I Ibragimov. "Independent and stationary sequences of random variables". In: *Wolters, Noordhoff Pub.* (1975).

[64] Edward Ignall and Arthur F Veinott Jr. "Optimality of myopic inventory policies for several substitute products". In: *Management Science* 15.5 (1969), pp. 284–304.

[65] Tetsuo Iida and Paul H Zipkin. "Approximate solutions of a dynamic forecast-inventory model". In: *Manufacturing & Service Operations Management* 8.4 (2006), pp. 407–425.

[66] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[67] Hao Jiang and Uday V Shanbhag. "On the solution of stochastic optimization and variational problems in imperfect information regimes". In: *SIAM Journal on Optimization* 26.4 (2016), pp. 2394–2429.

[68] Hao Jiang and Uday V Shanbhag. "On the solution of stochastic optimization problems in imperfect information regimes". In: *2013 Winter Simulations Conference (WSC)*. IEEE. 2013, pp. 821–832.

[69] Ruiwei Jiang and Yongpei Guan. "Data-driven chance constrained stochastic program". In: *Mathematical Programming* 158.1-2 (2016), pp. 291–327.

[70] Nathan Kallus and Xiaojie Mao. "Stochastic optimization forests". In: *arXiv preprint arXiv:2008.07473* (2020).

[71] L.V. Kantorovich and G.S. Rubinshtein. "On a space of totally additive functions". In: *Vestn.Leningr.Univ.* 13.52-59 (1958).

[72] Yi-hao Kao, Benjamin Roy, and Xiang Yan. "Directed regression". In: *Advances in Neural Information Processing Systems* 22 (2009), pp. 889–897.

[73] Robert S Kaplan. "A dynamic inventory model with stochastic lead times". In: *Management Science* 16.7 (1970), pp. 491–507.

[74] Kenji Kawaguchi. "Deep Learning without Poor Local Minima". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 586–594. URL: http://papers.nips.cc/paper/6112-deep-learning-without-poor-local-minima.pdf.

[75] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in Neural Information Processing Systems*. 2017, pp. 3146–3154.

[76] N Bora Keskin, Xu Min, and Jing-Sheng Jeannette Song. "The Nonstationary Newsvendor: Data-Driven Nonparametric Learning". In: *Available at SSRN 3866171* (2021).

[77] Rüdiger Kiesel. *Foundations of Modern Probability: Probability and Its Applications*. 1999.

[78] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[79] Danijel Kivaranovic, Kory D Johnson, and Hannes Leeb. "Adaptive, distribution-free prediction intervals for deep neural networks". In: *arXiv preprint arXiv:1905.10634* (2019).

[80] R. Koenker, A. Chesher, and M. Jackson. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005. ISBN: 9780521608275. URL: https://books.google.com/books?id=hdkt7V4NXsgC.

[81] Roger Koenker and Gilbert Bassett Jr. "Regression quantiles". In: *Econometrica: journal of the Econometric Society* (1978), pp. 33–50.

[82] Roger Koenker and Kevin F Hallock. "Quantile regression". In: *Journal of economic perspectives* 15.4 (2001), pp. 143–156.

[83] Ivana Komunjer. "Quantile Prediction, Chapter 17 in Handbook of Economic Forecasting (Edited by Elliott, G. and A. Timmermann). Volume 2, Chapter 17, 961-994". In: (2013).

[84] Leonid Aryeh Kontorovich, Kavita Ramanan, et al. "Concentration inequalities for dependent random variables via the martingale method". In: *The Annals of Probability* 36.6 (2008), pp. 2126–2158.

[85] James Kotary et al. "End-to-End Constrained Optimization Learning: A Survey". In: *arXiv preprint arXiv:2103.16378* (2021).

[86] Daniel Kuhn et al. "Wasserstein distributionally robust optimization: Theory and applications in machine learning". In: *Operations Research & Management Science in the Age of Analytics*. INFORMS, 2019, pp. 130–166.

[87] Vitaly Kuznetsov and Mehryar Mohri. "Generalization bounds for time series prediction with non-stationary processes". In: *International conference on algorithmic learning theory*. Springer. 2014, pp. 260–274.

[88] Jean Bernard Lasserre. *An introduction to polynomial and semi-algebraic optimization*. Vol. 52. Cambridge University Press, 2015.

[89] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), p. 436.

[90] Retsef Levi, Georgia Perakis, and Joline Uichanco. "The data-driven newsvendor problem: new bounds and insights". In: *Operations Research* 63.6 (2015), pp. 1294–1306.

[91] Retsef Levi, Robin O Roundy, and David B Shmoys. "Provably near-optimal sampling-based policies for stochastic inventory control models". In: *Mathematics of Operations Research* 32.4 (2007), pp. 821–839.

[92] Retsef Levi et al. "Approximation algorithms for stochastic inventory control models". In: *Mathematics of Operations Research* 32.2 (2007), pp. 284–302.

[93] Sergey Levine et al. "End-to-end training of deep visuomotor policies". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1334–1373.

[94] Jun Lin and Tsan Sheng Ng. "Robust multi-market newsvendor models with interval demand data". In: *European Journal of Operational Research* 212.2 (2011), pp. 361–373.

[95] Heyuan Liu and Paul Grigas. "Risk Bounds and Calibration for a Smart Predict-then-Optimize Method". In: *arXiv preprint arXiv:2108.08887* (2021).

[96] Liwan H Liyanage and J George Shanthikumar. "A practical inventory control policy using operational statistics". In: *Operations Research Letters* 33.4 (2005), pp. 341–348.

[97] Jayanta Mandi and Tias Guns. "Interior Point Solving for LP-based prediction+ optimisation". In: *arXiv preprint arXiv:2010.13943* (2020).

[98] Jayanta Mandi, Peter J Stuckey, Tias Guns, et al. "Smart predict-and-optimize for hard combinatorial optimization problems". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 02. 2020, pp. 1603–1610.

[99] Andreas Maurer. "A vector-contraction inequality for rademacher complexities". In: *International Conference on Algorithmic Learning Theory*. Springer. 2016, pp. 3–17.

[100] Daniel Mcdonald, Cosma Shalizi, and Mark Schervish. "Estimating beta-mixing coefficients". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. 2011, pp. 516–524.

[101] Daniel J McDonald, Cosma Rohilla Shalizi, and Mark Schervish. "Generalization error bounds for stationary autoregressive models". In: *arXiv preprint arXiv:1103.0942* (2011).

[102] A Yu Mitrophanov. "Sensitivity and convergence of uniformly ergodic Markov chains". In: *Journal of Applied Probability* 42.4 (2005), pp. 1003–1014.

[103] Mehryar Mohri and Afshin Rostamizadeh. "Stability bounds for non-iid processes". In: *Advances in Neural Information Processing Systems*. 2008, pp. 1025–1032.

[104] Mehryar Mohri and Afshin Rostamizadeh. "Stability bounds for stationary φ-mixing and β-mixing processes". In: *Journal of Machine Learning Research* 11.Feb (2010), pp. 789–814.

[105] Alp Muharremoglu and John N Tsitsiklis. "A single-unit decomposition approach to multiechelon inventory systems". In: *Operations Research* 56.5 (2008), pp. 1089–1103.

[106] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2003.

[107] Nam Ho-Nguyen and Fatma Kılınç-Karzan. "Exploiting problem structure in optimization under uncertainty via online convex optimization". In: *Mathematical Programming* 177.1 (2019), pp. 113–147.

[108] Nam Ho-Nguyen and Fatma Kılınç-Karzan. "Risk Guarantees for End-to-End Prediction and Optimization Processes". In: *arXiv preprint arXiv:2012.15046* (2020).

[109] Afshin Oroojlooyjadid, Lawrence Snyder, and Martin Takáč. "Applying deep learning to the newsvendor problem". In: *arXiv preprint arXiv:1607.02177* (2016).

[110] Afshin Oroojlooyjadid, Lawrence V Snyder, and Martin Takáč. "Applying deep learning to the newsvendor problem". In: *IISE Transactions* 52.4 (2020), pp. 444–463.

[111] Adam Paszke et al. "Automatic differentiation in PyTorch". In: (2017).

[112] Georgia Perakis and Guillaume Roels. "Regret in the newsvendor model with partial information". In: *Operations Research* 56.1 (2008), pp. 188–203.

[113] Marin Vlastelica Pogančić et al. "Differentiation of blackbox combinatorial solvers". In: *International Conference on Learning Representations*. 2019.

[114] Meng Qi, Ying Cao, and Zuo-Jun Shen. "Distributionally robust conditional quantile prediction with fixed design". In: *Management Science* (2021).

[115] Meng Qi, Ying Cao, and Zuo-Jun Shen. "Distributionally robust conditional quantile prediction with fixed design". In: *Management Science* 68.3 (2022), pp. 1639–1658.

[116] Meng Qi, Zuo-Jun Max Shen, and Zeyu Zheng. "Learning newsvendor problem with intertemporal dependence and moderate non-stationarities". In: *Available at SSRN 3648615* (2020).

[117] Meng Qi et al. "A practical end-to-end inventory management model with deep learning". In: *Available at SSRN 3737780* (2020).

[118] Ruozhen Qiu, Jennifer Shang, and Xiaoyuan Huang. "Robust inventory decision under distribution uncertainty: A CVaR-based optimization approach". In: *International Journal of Production Economics* 153 (2014), pp. 13–23.

[119] Hamed Rahimian, Güzin Bayraksan, and Tito Homem-de-Mello. "Distributionally Robust Newsvendor Problems with Variation Distance". In: *Available at Optimization Online* (2017).

[120] Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. "Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes". In: *Journal of Machine Learning Research* 11.Jul (2010), pp. 1927–1956.

[121] Vivek Ramamurthy, J George Shanthikumar, and Zuo-Jun Max Shen. "Inventory policy with parametric demand: Operational statistics, linear correction, and regression". In: *Production and Operations Management* 21.2 (2012), pp. 291–308.

[122] Yaniv Romano, Evan Patterson, and Emmanuel Candes. "Conformalized quantile regression". In: *Advances in Neural Information Processing Systems*. 2019, pp. 3538–3548.

[123] Paul R Rosenbaum and Donald B Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (1983), pp. 41–55.

[124] Saharon Rosset and Ryan J Tibshirani. "From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation". In: *Journal of the American Statistical Association* (2018), pp. 1–14.

[125] Donald B Rubin and Richard P Waterman. "Estimating the causal effects of marketing interventions using propensity score methodology". In: *Statistical Science* (2006), pp. 206–222.

[126] Herman Rubin. "Uniform convergence of random functions with applications to statistics". In: *The Annals of Mathematical Statistics* (1956), pp. 200–203.

[127] Anna-Lena Sachs. "The data-driven newsvendor with censored demand observations". In: *Retail Analytics*. Springer, 2015, pp. 35–56.

[128] Paul-Marie Samson. "Concentration of measure inequalities for Markov chains and ϕ-mixing processes". In: *The Annals of Probability* 28.1 (2000), pp. 416–461.

[129] Herbert Scarf. "A min-max solution of an inventory problem". In: *Studies in the mathematical theory of inventory and production* (1958).

[130] Chuen-Teck See and Melvyn Sim. "Robust approximation to multiperiod inventory management". In: *Operations research* 58.3 (2010), pp. 583–594.

[131] Sanjit A Seshia et al. "Formal specification for deep neural networks". In: *International Symposium on Automated Technology for Verification and Analysis*. Springer. 2018, pp. 20–34.

[132] Glenn Shafer and Vladimir Vovk. "A tutorial on conformal prediction". In: *Journal of Machine Learning Research* 9.Mar (2008), pp. 371–421.

[133] Simon J Sheather and Michael C Jones. "A reliable data-based bandwidth selection method for kernel density estimation". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.3 (1991), pp. 683–690.

[134] L.V. Snyder and Z.J.M. Shen. *Fundamentals of Supply Chain Theory*. Wiley, 2011. ISBN: 9780470521304. URL: https://books.google.com/books?id=U7GTrLyVnPMC.

[135] Peter Sprent and Nigel C Smeeton. *Applied nonparametric statistical methods*. CRC press, 2016.

[136] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958.

[137] Masashi Sugiyama and Amos J Storkey. "Mixture regression for covariate shift". In: *Advances in Neural Information Processing Systems*. 2007, pp. 1337–1344.

[138] Masashi Sugiyama, Makoto Yamada, and Marthinus Christoffel du Plessis. "Learning under nonstationarity: covariate shift and class-balance change". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 5.6 (2013), pp. 465–477.

[139] Rangarajan K Sundaram et al. *A first course in optimization theory*. Cambridge university press, 1996.

[140] Ryan J Tibshirani et al. "Conformal Prediction Under Covariate Shift". In: *Advances in Neural Information Processing Systems*. 2019, pp. 2526–2536.

[141] L Beril Toktay and Lawrence M Wein. "Analysis of a forecasting-production-inventory system with stationary demand". In: *Management Science* 47.9 (2001), pp. 1268–1281.

[142] Arthur F Veinott Jr. "Optimal policy for a multi-product, dynamic, nonstationary inventory problem". In: *Management Science* 12.3 (1965), pp. 206–222.

[143] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.

[144] Kai Wang, Boris Babenko, and Serge Belongie. "End-to-end scene text recognition". In: *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE. 2011, pp. 1457–1464.

[145] Tong Wang, Atalay Atasu, and Mümin Kurtuluş. "A multiordering newsvendor model with dynamic forecast evolution". In: *Manufacturing & Service Operations Management* 14.3 (2012), pp. 472–484.

[146] Zizhuo Wang, Peter W Glynn, and Yinyu Ye. "Likelihood robust optimization for data-driven newsvendor problems". In: *Preprint* (2009).

[147] Zizhuo Wang, Peter W Glynn, and Yinyu Ye. "Likelihood robust optimization for data-driven problems". In: *Computational Management Science* 13.2 (2016), pp. 241–261.

[148] Ruofeng Wen, Kari Torkkola, and Balakrishnan Narayanaswamy. "A Multi-Horizon Quantile Recurrent Forecaster". In: *arXiv preprint arXiv:1711.11053* (2017).

[149] Bryan Wilder, Bistra Dilkina, and Milind Tambe. "Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 1658–1665.

[150] Bryan Wilder et al. "End to end learning and optimization on graphs". In: *arXiv preprint arXiv:1905.13732* (2019).

[151] Adam E Wyse and Dong Gi Seo. "A Comparison of Three Conditional Growth Percentile Methods: Student Growth Percentiles, Percentile Rank Residuals, and a Matching Method." In: *Practical Assessment, Research & Evaluation* 19 (2014).

[152] Jia Zhai, Hui Yu, and Caihong Sun. "Robust optimization for the newsvendor problem with discrete demand". In: *Mathematical Problems in Engineering* 2018 (2018).

[153] Chaoyue Zhao and Yongpei Guan. "Data-driven risk-averse stochastic optimization with Wasserstein metric". In: *Operations Research Letters* 46.2 (2018), pp. 262–267.

[154] Kaijie Zhu and Ulrich W Thonemann. "An adaptive forecasting algorithm and inventory policy for products with short life cycles". In: *Naval Research Logistics (NRL)* 51.5 (2004), pp. 633–653.

[155] Zhisu Zhu, Jiawei Zhang, and Yinyu Ye. "Newsvendor optimization with limited distribution information". In: *Optimization methods and software* 28.3 (2013), pp. 640–667.

[156] P. Zipkin. *Foundations of Inventory Management*. McGraw-Hill Companies,Incorporated, 2000. ISBN: 9780256113792. URL: https://books.google.com/books?id=rjzbkQEACAAJ.

Appendix A

Supplementary material for Chapter 2

A.1 Proof of Theorem 2.3.1

In order to prove Theorem 1, we need the following lemma.

Lemma A.1.1. *Consider the total inventory cost within time period* $[t_1, t_2]$ *as* $f(a) = \sum_{s=t_1}^{t_2} h[I_{t_1} + a - d_{[t_1,s]}]^+ + b[d_{[t_1,s]} - I_{t_1} - a]^+$, *where* h *is the unit holding cost and* b *is the unit back-order cost. The optimal order quantity which minimizes the total cost is derived by* $\arg\min_{a\in\mathbb{R}} f(a) = a^* = d_{[t_1,s^*]} - I_{t_1}$ *where* $s^* = \lfloor \frac{b(t_2+1-t_1)}{h+b} + t_1 \rfloor$.

Proof. Proof of Lemma A.1.1: First we show that a^* takes the form of $d_{[t_1,s]} - I_{t_1}$, for some $s \in [t_1, t_2]$. It is evident that $f(a)$ is convex and piece-wise linear, hence the optimal solution should be one of the extreme point, which is $d_{[t_1,s]} - I_{t_1}$ for some s. Now we need to choose s^* such that $s^* = \arg\min_{s\in\{t_1,\ldots,t_2\}}\{f(d_{[t_1,s]} - I_{t_1})\}$. Note that for $s \in \{t_1, \ldots, t_2\}$, choosing to satisfy one unit of demand that occurred at period s generates a holding cost $h(s - t_1)$, while choosing not to satisfy one unit of demand generates back-order cost $b(t_2 + 1 - s)$, we will choose to satisfy the demand unit that occurs at period s such that $h(s - t_1) \leq b(t_2 + 1 - s) = b(t_2 + 1 - t_1) - b(s - t_1)$. The above analysis is valid for all units that occurs in period s, hence we will choose to satisfy either all demand units that occurs at period s or non of them. Hence we have $s \leq s^*$ for all s such that $h(s-t_1) \leq b(t_2+1-t_1)-b(s-t_1)$, which lead to the final solution $s^* = \lfloor \frac{b(t_2+1-t_1)}{h+b} + t_1 \rfloor$. □

Lemma A.1.1 leads to our final result of Theorem 1, on the optimal order quantity under the realization of each sample data.

Proof. Proof of Theorem 2.3.1: First, we relax the constraint $a_m \geq 0$ and start from the last decision a_M. By Lemma A.1.1, $a_M^* = \min_{s\in\{v_M,\ldots,T\}}\{d_{[v_M,s]} - I_{v_M}\}$ and $s^* = \arg\min_{s'\in\{v_M,\ldots,T\}} \sum_{s=v_M}^{T} h[d_{v_M,s'} - d_{[v_M,s]}]^+ + b[d_{[v_M,s]} - d_{v_M,s'}]^+$. It should be noted that s^*

only depends on the sequence $\{d_t\}_{t=v_M}^{T}$. Hence, $f_M^* := f(a_M^*) = \sum_{s=v_M}^{T} h[d_{v_M,s^*} - d_{[v_M,s]}]^+ + b[d_{[v_M,s]} - d_{v_M,s^*}]^+$ does not depend on I_{v_M}; hence, the decision of a_M^* is independent to the decision of a_{M-1}^*. By recursion, we can decompose the decisions $\{a_m^*\}_{m=1}^{M}$ with corresponding period $[v_m, v_{m+1})$.

Now we consider the constraint $a_m \geq 0$. It is not difficult to verify that with constraint $a_m \geq 0$, we can still compute a_m^* use $a_m^* = \arg\min_{a_m \geq 0} \sum_{s=v_m}^{v_{m+1}-1} h[I_{v_m} + a_m - d_{[v_m,s]}]^+ + b[d_{[v_m,s]} - I_{v_m} - a_m]^+$ in the sequence of $a_1, a_2, \ldots, a_M$. We let a_m^{**} denote the optimal solution with constraint $a_m \geq 0$. Note that under this constraint, f_{m+1}^* depends on $I_{v_{m+1}}$. In fact, it is a non-decreasing function of $I_{v_{m+1}}$, hence, a non-decreasing function of a_m. Suppose for m, following the above method, we have $a_m^* < 0$, then $a_m^{**} = 0$ is not only the minimizer of $f_m(a_m)$ but also a minimizer of f_{m+1}^*. □

A.2 Sensitivity Analysis

In this part, we conduct various sensitivity analysis to demonstrate the robustness and generalization ability of proposed E2E model under different hyper-parameter choices, data size and model covariates.

A.2.1 Sensitivity Analysis for Network Hyper-parameters

We first provide details of the neural network structure and hyper-parameters used in the E2E model. In the training stage, we sweep through different combinations of hyper-parameters within the considered range and use the validation set to choose the best hyper-parameter values, which is shown below as the "Default Value" column.

Hyperparameter	Default Value	Range
`DF_submodule: MQRNN`	Hidden state size 50	{30, 40, 50, 60}
`VLT_submodule` (`VLT_layer_1`, `VLT_layer_2`)	Layer size {50, 20}	{100, 20}, {50, 20}, {30, 20}
Integration module (`Layer_3`, `Layer_4`)	Layer size {100, 100}	{100}, {100, 100}, {100, 100, 100}, {100, 100, 100, 100}
Learning rate	0.001	{0.0001, 0.001, 0.01}
Learning rate decay	$1e-4$	{1e-5, 1e-4, 1e-3}
Momentum	0.85	{0.8, 0.85, 0.9}
Mini-batch size	64	{64, 128, 256}
Weight initialization	Gaussian $\mu = 0, \sigma = 0.01$	{Gaussian, Uniform}
Activation	Rectified linear unit (ReLU)	{ReLU, tanh}
Dropout rate	0.2	{0.1, 0.2, 0.3, 0.4}

Table A.1: Network hyper-parameters

In addition, in order to investigate the sensitivity of the E2E model performance stated

in Section 2.4 with respect to the network structure and hyper-parameters, we provide three sets of sensitivity tests with respect to the following hyper-parameters:

- Number of hidden layers
- Number of neurons/weights
- Learning rate

A.2.1.1 Number of hidden layers:

To check the sensitivity of E2E model with respect to the number of hidden layers, we adjust the number of hidden layers in the integration module. By default, there are two hidden layers `Layer_3` and `Layer_4` both of size 100. We tried three other variants: 1) keep only one hidden layer, i.e., `Layer_3` of size 100; 2) have three hidden layers, i.e., `Layer_3`, `Layer_4`, `Layer_5`, each of size 100; 4) have four hidden layers, i.e., `Layer_3`, `Layer_4`, `Layer_5`, `Layer_6`, each of size 100. We trained four different E2E models with the four different network structures, and test their performance using the same dataset and experiment setup as in Section 4.1.

	OPT	E2E (1 layer)	E2E (2 layers)	E2E (3 layers)	E2E (4 layers)
Total cost	3022.60	**4857.20 (+60.1%)**	**3766.52(+24.6%)**	**3822.69 (+26.5%)**	**3904.74(+29.2%)**
Holding cost	1925.70	3919.17	2689.02	2453.32	2231.01
Stockout cost	1096.91	938.03	1077.50	1369.38	1673.73

Table A.2: Sensitivity of E2E model w.r.t. to the number of hidden layers

Table A.2 shows the total inventory management cost, holding cost and stockout cost of the four E2E models with different number of hidden layers. When the hidden layer number is 1, the end-to-end model performance is inferior to other benchmark models (in Figure 2) due to the lack of representation capacity. In all other cases, the end-to-end models outperform the benchmarks. As the number of hidden layers increases, the cost of end-to-end model first decreases then increases. By having more layers, the network representation power increases and can better fit the relationship between observation and the optimal order decision. However, an over-complicated network may cause over-fitting in the training set and shows less generalization capacity in test set.

A.2.1.2 Number of neurons/weights:

To investigate the model sensitivity w.r.t. the number of neurons, we sweep through $[30, 40, 50, 60]$ as the hidden state size of the demand prediction modile (MQRNN), and compare the performance of the different networks.

	OPT	E2E (30)	E2E (40)	E2E (50)	E2E(60)
Total cost	3022.60	**3846.28(+27.3%)**	**3783.32(+25.2%)**	**3766.52(+24.6%)**	**3860.99(+27.7%)**
Holding cost	1925.70	2886.45	2710.71	2689.02	2185.18
Stockout cost	1096.91	959.83	1072.61	1077.50	1675.81

Table A.3: Sensitivity of E2E model w.r.t. to the number of neurons

Table A.3 reports the total cost, holding cost and stockout cost of the four E2E models with different number of neurons. As expected, as the number of hidden neurons increases, the E2E model cost first decreases then increases. Similar as the effect of adding more layers, by adding more neurons, the representation capacity of the neural network enhances. However, too many neurons may lead to over-fitting in the training set and the trained network shows worse generalization performance.

A.2.1.3 Learning rate:

Learning rate is one of the most important hyper-parameter in neural network training [89]. Table A.4 shows how different learning rate value affects the performance of E2E model on the test set. If we further increase the learning rate to 0.1 or larger, network training process becomes unstable and the training loss oscillates which leads to much worse performance. When learning rate is 0.0001, the learnt model is slightly better than our default setting ($lr = 0.001$) but the training time increases significantly.

	OPT	E2E (lr =0.0001)	E2E (lr =0.001)	E2E (lr =0.01)
Total cost	3022.60	**3744.28 (+23.9%)**	**3766.52(+24.6%)**	**4231.01(+39.9%)**
Holding cost	1925.70	2424.71	2689.02	2350.98
Stockout cost	1096.91	1319.58	1077.50	1880.03

Table A.4: Sensitivity w.r.t. to learning rate

A.2.2 Sensitivity Analysis for Data Size

In order to investigate the sensitivity w.r.t. training data size, we use $[20\%, 40\%, 60\%, 80\%, 100\%]$ of the training data to train the end-to-end model. The cost and computational time of the E2E model with different amounts of training data are provided in Table A.5 and Figure A.1.

A.2.3 Sensitivity Analysis for b and h

All the previous simulations are based on the assumption that the ratio between unit stock-out cost and unit holding cost is 9, which may not reflect the case in real world.

Percentage	Number of training data	OPT cost	E2E cost	E2E Training time (s)
20%	3893 SKUs	3022.60	5608.01(+85.5%)	281.61
40%	7786 SKUs	3022.60	4765.52(+57.7%)	624.64
60%	11680 SKUs	3022.60	4226.84(+39.8%)	1022.57
80%	15573 SKUs	3022.60	3783.54(+25.2%)	1459.42
100%	19466 SKUs	3022.60	3766.52(+24.6%)	1830.36

Table A.5: E2E Model Sensitivity w.r.t. to Training Data Size

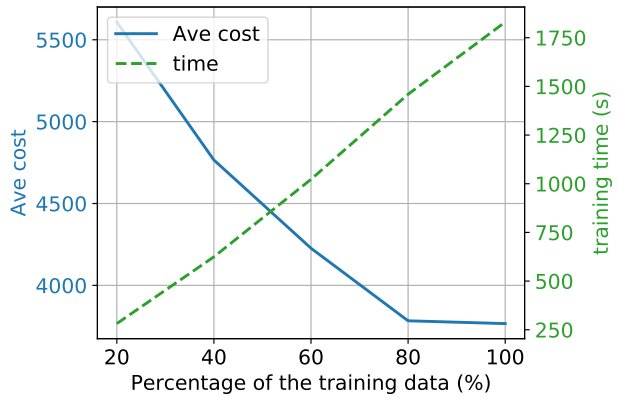


Figure A.1: E2E Model Sensitivity w.r.t. to Training Data Size

Therefore, sensitivity analysis with respect to different values of b/h is conducted below to validate our conclusion in previous parts. It should be noted that different b/h ratios not only lead to different measurements costs, but also change replenishment decisions for all models.

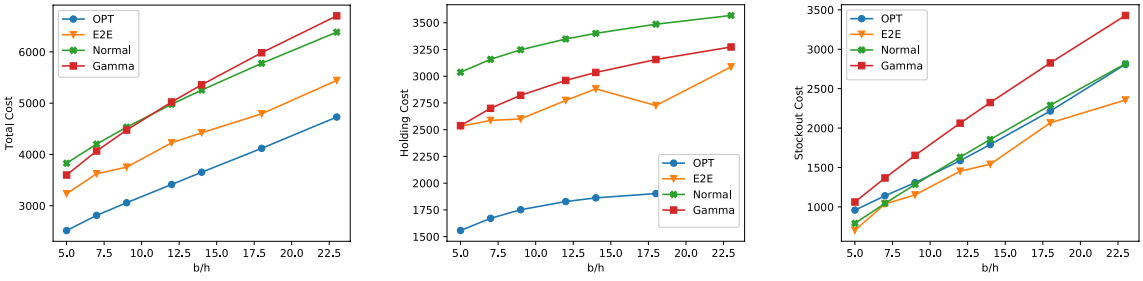


Figure A.2: Performance of each model with different ratios on b/h.

Figure A.2 demonstrates total cost, holding cost as well as stockout cost of E2E model,

OPT decision and two base stock algorithms under different b/h ratios. According to these results, E2E model is stable and closest to the optimal solutions in most cases, with respect to different choices of b/h.

A.2.4 Sensitivity for Different Neural Network Local Minima

Since the loss function of neural networks are highly non-convex and contain many local minima, when training neural networks using stochastic gradient descent (SGD) methods (the classical way of training neural networks), we are very likely to get stuck in one of the local minima. However, several recent research experimenting with larger networks and SGD suggest that, while deep neural networks do have many local minima, they consistently give very similar performance [35, 74]. Consistent with the above results from machine learning literature, we find the performance of our proposed E2E deep learning model has similar property in numerical experiments. For instance, two different E2E models with the default network structure and hyper-parameter are trained with different weight initialization (random seeds). Figure A.3 visualizes the final weights of `VLT_layer_1` of the two trained networks. And Table A.6 provides the performances of two networks on test data set. It can be observed that although the two network weights are quite different, they have similar performances on the test set in terms of total cost, holding cost as well as stockout cost.

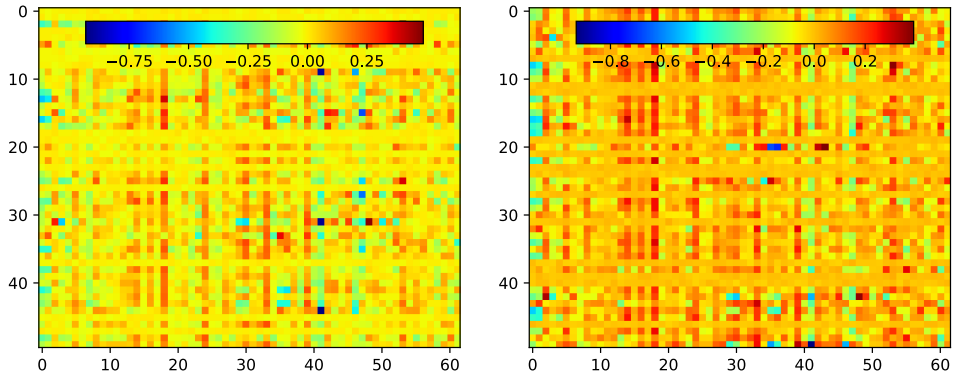


Figure A.3: Network sensitivity w.r.t. local minima. The two plots show the weights of `VLT_layer1` of two neural networks trained with different initialization.

A.2.5 Local Explainability for E2E Model

Finally, we exam the E2E model performance with respective to some important model covariates, e.g., initial inventory level and review period. In particular, we check the network performance regarding some formal specifications [131] to ensure the network performance is aligned with the common sense for inventory management. For example, if the initial

	OPT	E2E (Network 1)	E2E (Network 2)
Total cost	3022.60	3766.52(+24.6%)	3800.54(+25.7%)
Holding cost	1925.70	2689.02(+39.6%)	2766.67(+43.7%)
Stockout cost	1096.91	1077.50(-1.8%)	1033.87(-5.7%)

Table A.6: Comparison of two E2E model performance with different initialization seeds.

inventory level at time t_1 is strictly greater than time t_2 with all other covariates being same, then order amount at t_1 should be lower than t_2. Similar for the review period, the order amount should be higher when the the review period is longer.

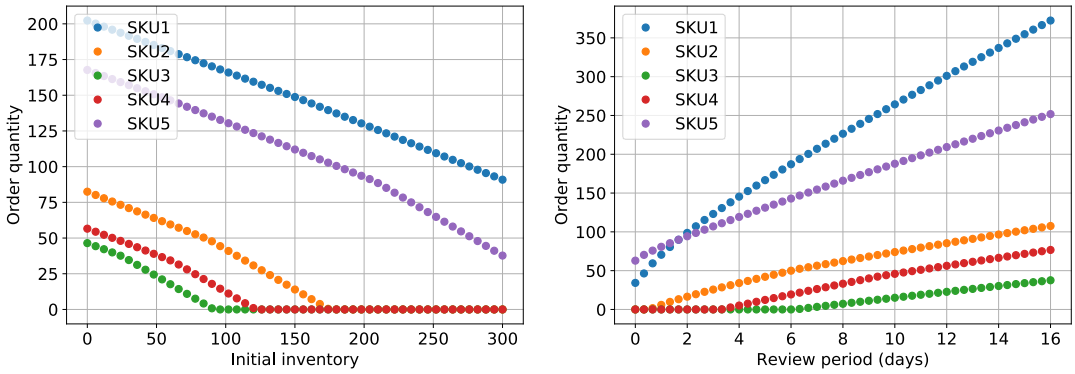


Figure A.4: Order quantity as a function of covariates of different input features

Figure A.4 visualizes the one-dimensional relationship between the E2E model output (order quantity), initial inventory level and review period for 5 different SKUs. It should be noted that the same trend holds for all SKUs, while the slope and shape might be different for different SKUs. Notably, the sensitivity results of the end-to-end model with respective to both initial inventory level and review period satisfy the monotonicity specification. As the initial inventory level increases, the order quantity monotonically decreases. Once the initial inventory reaches certain threshold, the order quantity reduces to 0 as there's enough inventory before next review cycle and no need for further ordering. Besides, the model output increases monotonically with respect to review period. The longer the review period is, the more the order quantity should be.

Appendix B

Supplementary material for Chapter 3

B.1 Supplementary Lemmas and Proofs

B.1.1 Lemmas and Proofs for Section 3.3

Lemma B.1.1. *For any $\rho > 0$, $w_\rho(\cdot)$ is a single-valued continuous function of $p \in \Delta_K$.*

Proof. Proof of Lemma B.1.1 From part 2 in Theorem 9.17 of [139], the mapping $w(p) : \Delta_k \to S$ is a continuous function of p. □

Lemma B.1.2. *Suppose that Assumptions 1.A and 1.B hold, and let $w(\cdot) : \Delta_K \to S$ be an optimal solution mapping such that $w(p) \in W(p)$ for all $p \in \Delta_K$. Then, we have that: (i) f^* is the unique minimizer of $\min_{f\in\mathcal{H}} R(w \circ f; 0)$; (ii) the optimal expected unregularized risk is $J^* := \min_{f\in\mathcal{H}} R(w \circ f; 0) = R(w \circ f^*, 0)$, and the value of J^* is the same for all valid optimal solution mappings $w(\cdot)$.*

Proof. Proof of Lemma B.1.2 We first prove part *(i)*. Suppose that there exists $\bar{f} \neq f^*$ in $\mathcal{H}$ such that

$$\mathbb{E}_x \left[\sum_{k=1}^{K} p_k^*(x) c_k(w(\bar{f}(x))) \right] \leq \mathbb{E}_x \left[\sum_{k=1}^{K} p_k^*(x) c_k(w(f^*(x))) \right].$$

Since $w(f^*(x)) \in W(f^*(x))$ and $f^*(x)$ is the true hypothesis, i.e., $f_k^*(x) = p_k^*(x)$, we have

$$\sum_{k=1}^{K} p_k^*(x) c_k(w(\bar{f}(x))) \geq \sum_{k=1}^{K} p_k^*(x) c_k(w(f^*(x)))$$

for all $x \in \mathcal{X}$. Combining the above two inequalities yields

$$\sum_{k=1}^{K} p_k^*(x) c_k(w(\bar{f}(x))) = \sum_{k=1}^{K} p_k^*(x) c_k(w(f^*(x))),$$

almost surely for all $x \in \mathcal{X}$. Hence, $w(\bar{f}(x)) \in W(f^*(x))$ almost surely, which contradicts Assumption 1.B. Therefore, under Assumptions 1.A and 1.B, f^* is the unique minimizer of $\min_{f \in \mathcal{H}} R(w \circ f; 0)$.

We now show *(ii)*. Let $w(\cdot), u(\cdot) : \Delta_K \to S$ be two valid optimal solution mappings such that $w(p) \in W(p)$ and $u(p) \in W(p)$ for all $p \in \Delta_K$. Thus, $w(f^*(x)) \in W(f^*(x))$ and $u(f^*(x)) \in W(f^*(x))$ for all x, which yields $\sum_{k=1}^{K} p_k^*(x) c_k(w(f^*(x))) = \sum_{k=1}^{K} p_k^*(x) c_k(u(f^*(x)))$ by the definition of $W(\cdot)$. Therefore, taking expectation with respect to x yields $J^* = R(w \circ f^*, 0) = R(u \circ f^*, 0)$. □

Proof. Proof of Corollary 3.3.3.1 According to Theorem 3.3.3, we have

$$R(w_{\rho_n} \circ \hat{f}^n_{\rho_n}) \;\leq\; \hat{R}_n(w_{\rho_n} \circ \hat{f}^n_{\rho_n}; \rho_n) + \frac{\sqrt{2} L_c^2}{\rho_n} \mathfrak{R}_n(\mathcal{H}) + \bar{c} \sqrt{\frac{\log(\frac{2}{\delta})}{2n}}$$

with probability at least $1 - \frac{\delta}{2}$. Since $\hat{f}^n_{\rho_n}$ is the ICEO minimizer,

$$\hat{R}_n(w_{\rho_n} \circ \hat{f}^n_{\rho_n}; \rho_n) \leq \hat{R}_n(w_{\rho_n} \circ f^*; \rho_n) \leq \hat{R}_n(w_{\rho_n} \circ f^*) + \rho_n \bar{\phi}.$$

Then we apply Hoeffding's inequality and have

$$\hat{R}_n(w_{\rho_n} \circ f^*) \leq R_n(w_{\rho_n} \circ f^*) + \bar{c} \sqrt{\frac{2 \log(\frac{2}{\delta})}{n}}$$

with probability at least $1 - \frac{\delta}{2}$. Combining the above two inequalities leads to the desired result. □

B.1.2 Proofs for Section 3.4

Proof. Proof of Proposition 3.4.1.2 Considering the kernel function $\mathcal{K}(p, p') = (c + p^T p')^s$ with $p \in \mathbb{R}^K$, we first generalize the result of Example 13.19 from [143]. When the input p and p' are K-dimensional vectors, the empirical kernel matrix can have rank at most $\frac{(s-1+K)!}{(s-1)!K!}$. Therefore, the left-hand side of Inequality (13.56) from [143] can be upper-bounded by $\delta_m \sqrt{\frac{1}{m} \frac{(s-1+K)!}{(s-1)!K!}}$. Then we can apply Theorem 13.17 from [143] and set $\lambda_m = 2\delta_m^2$ to achieve inequality (3.4.1.2). Moreover, the empirical Rademacher complexity can be upper-bounded by $\bar{c} \sqrt{\frac{1}{m} \frac{(s-1+K)!}{(s-1)!K!}}$ with some constant $\bar{c}$. Then if we have $\theta_m \geq \bar{c}_3 b \sqrt{\frac{1}{m} \frac{(s-1+K)!}{(s-1)!K!}}$, we can apply Theorem 14.1 from [143] and therefore have the desired result inequality (3.4.1.2). □

Proof. Proof of Corollary 3.4.1.1 The proof follow from a slight modification of the proof of Theorem 3.4.1. We first consider

$$\frac{1}{n}\sum_{i=1}^{n}\|w_\rho(f(x_i))) - \tilde{w}_\rho(f(x_i))\|_1 \leq L_c \sum_{j=1}^{d}(\frac{1}{n}\sum_{i=1}^{n}|w_{\rho,j}(f(x_i)) - \tilde{w}_{\rho,j}(f(x_i))|^2)^{\frac{1}{2}}.$$

Then noted that

$$\mathbb{E}_{\mathcal{D}_{f(x)}}[|w_{\rho,j}(p)) - \tilde{w}_{\rho,j}(p)|^2] \leq \mathbb{E}_{\mathcal{D}_p}[|w_{\rho,j}(p)) - \tilde{w}_{\rho,j}(p)|^2] + 2\bar{w}_j^2 \mathrm{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p),$$

because $|w_{\rho,j}(p)) - \tilde{w}_{\rho,j}(p)|^2$ is bounded by $\bar{w}_j^2$ for all $p \in \Delta_K$. Thus,

$$\mathbb{E}_{\mathcal{D}_{f(x)}}[|w_{\rho,j}(p)) - \tilde{w}_{\rho,j}(p)|^2]^{1/2} \leq \mathbb{E}_{\mathcal{D}_p}[|w_{\rho,j}(p)) - \tilde{w}_{\rho,j}(p)|^2]^{1/2} + \bar{w}_j\sqrt{2\mathrm{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p)}.$$

Following the same reasoning of the proof in Theorem 3.4.1, the desired result follows. □

Proof. Proof of Theorem 3.4.1 By Theorem 3.3.3 (i), for any ρ, we have

$$R(w_\rho \circ f; \rho) \leq \hat{R}_n(w_\rho \circ f; \rho) + \frac{\sqrt{2}L_c^2}{\rho}\mathfrak{R}_n(\mathcal{H}) + \bar{c}\sqrt{\frac{\log(\frac{1}{\delta})}{2n}}.$$

Noted that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}|c(w_\rho(f(x_i)), \xi_i) - c(\tilde{w}_\rho(f(x_i)), \xi_i)| &\leq L_c\frac{1}{n}\sum_{i=1}^{n}\|w_\rho(f(x_i))) - \tilde{w}_\rho(f(x_i))\|_1 \\ &= L_c\sum_{j=1}^{d}\frac{1}{n}\sum_{i=1}^{n}|w_{\rho,j}(f(x_i)) - \tilde{w}_{\rho,j}(f(x_i))| \quad \text{(B.1)}\end{aligned}$$

When the approximated oracle has a uniform error, we combine (B.1) and Assumption 3 to have

$$\frac{1}{n}\sum_{i=1}^{n}|c(w_\rho(f(x_i)), \xi_i) - c(\tilde{w}_\rho(f(x_i)), \xi_i)| \leq L_c\sum_{j=1}^{d}\mathcal{E}_j^{\text{unif}}. \quad \text{(B.2)}$$

When the oracle is noised, we consider two different distributions $\mathcal{D}_{f(x)}$ and $\mathcal{D}_p$. We let $\mathcal{D}_{f(x)}$ denote the distribution of $f(x)$ given a hypothesis f and the distribution of x, $\mathcal{D}_x$. Moreover, we let $\mathcal{D}_p$ denote the distribution used to generate training samples $\{(p_i, w_i)\}_{i=1}^m$ for oracle approximation. Then, we apply the error bound (4) with distribution $\mathcal{D}_{f(x)}$ and $\mathcal{D}_p$ respectively and have

$$\frac{1}{n}\sum_{i=1}^{n}|w_{\rho,j}(f(x_i)) - \tilde{w}_{\rho,j}(f(x_i))| \leq \mathbb{E}_{\mathcal{D}_{f(x)}}[|w_{\rho,j}(p)) - \tilde{w}_{\rho,j}(p)|] + \mathcal{E}_j^{\text{prob}}(n, \delta/2d; \mathcal{G}),$$

and

$$\mathbb{E}_{\mathcal{D}_p}[|w_{\rho,j}(p)) - \tilde{w}_{\rho,j}(p)|] \leq \frac{1}{m}\sum_{i=1}^{m} |w_{\rho,j}(p_i)) - \tilde{w}_{\rho,j}(p_i)| + \mathcal{E}_j^{\text{prob}}(m, \delta/2d; \mathcal{G}),$$

each with probability at least $1 - \frac{\delta}{2d}$. Considering the total variation between $\mathcal{D}_{f(x)}$ and $\mathcal{D}_p$, we have the following

$$\mathbb{E}_{\mathcal{D}_{f(x)}}[|w_{\rho,j}(p)) - \tilde{w}_{\rho,j}(p)|] \leq \mathbb{E}_{\mathcal{D}_p}[|w_{\rho,j}(p)) - \tilde{w}_{\rho,j}(p)|] + \text{diam}_j(S)\text{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p),$$

where $\text{diam}_j(S)$ is the diameter of S in the j-th coordinate and TV denotes the total variation. This result holds because $|w_{\rho,j}(\cdot) - \tilde{w}_{\rho,j}(\cdot)|$ is continuous and bounded by $\text{diam}_j(S)$. Thus,

$$\frac{1}{n}\sum_{i=1}^{n} \|w_\rho(f(x_i))) - \tilde{w}_\rho(f(x_i))\|_1 \leq \sum_{j=1}^{d}\left[\frac{1}{m}\sum_{i=1}^{m} |w_{\rho,j}(p_i) - \tilde{w}_{\rho,j}(p_i)| + \mathcal{E}_j^{\text{prob}}(n, \delta/2d; \mathcal{G}) + \mathcal{E}_j^{\text{prob}}(m/2d, \delta; \mathcal{G})\right] + \text{diam}_j(S)\text{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p),$$

with probability at least $1 - \delta$. Therefore, we have

$$(\text{B.1}) \leq L_c \sum_{j=1}^{d}\left[\frac{1}{m}\sum_{i=1}^{m} |w_{i,j} - \tilde{w}_{\rho,j}(p_i)| + \mathcal{E}_j^{\text{prob}}(n, \delta/2d; \mathcal{G}) + \mathcal{E}_j^{\text{prob}}(m, \delta/2d; \mathcal{G})\right] + \text{diam}(S)L_c\text{TV}(\mathcal{D}_{f(x)}, \mathcal{D}_p), \tag{B.3}$$

with probability at least $1 - \delta$. Here we slightly abuse the notation and let $\text{diam}(S)$ denote the summation of coordinate-wise diameter along all coordinates.

Then (3.12) and (3.14) follow from combining (B.2) and (B.3) with Theorem 3.3.3. (3.13) and (3.15) follow from the non-negativity of the regularization term. □

Proof. Proof of Proposition 3.4.1.3: We first consider the constraint $B^T x + b \geq 0, \forall x, \text{s.t.} \quad Ax \geq a$ is equivalent to

$$0 \leq \min \quad B_k^T x + b_k \tag{B.4}$$
$$\text{s.t.} \quad Ax \geq a \tag{B.5}$$

for all $k = 1, \ldots, K$. Then consider the dual of (B.4)-(B.5)

$$-b_k \leq \max \quad a^T y_k$$
$$\text{s.t.} \quad A^T y_k = B_k$$
$$y_k \geq 0$$

which reduces to find a feasible solution of the following group of constraints

$$\begin{cases} a^T y_k \geq -b_k \\ A^T y_k = B_k \\ y_k \geq 0 \end{cases} \tag{B.6}$$

for all $k = 1, \ldots, K$.
Then the normalization constraint $\mathbb{1}^T(Bx + b) \geq 1, \forall x, \text{s.t.} Ax \geq a$ is equivalent to

$$1 \leq \min \quad \mathbb{1}^T Bx + \mathbb{1}^T b \tag{B.7}$$
$$\text{s.t.} \quad Ax \geq a, \tag{B.8}$$

similarly by considering the dual problem

$$\begin{aligned} 1 - \mathbb{1}^T b \leq \max \quad & a^T z \\ \text{s.t.} \quad & A^T z = B^T \mathbb{1} \\ & z \geq 0, \end{aligned}$$

which reduces to the following group of constraints

$$\begin{cases} a^T z \geq 1 - \mathbb{1}^T b \\ A^T z = B^T \mathbb{1} \\ z \geq 0. \end{cases} \tag{B.9}$$

Finally, the constraint $\mathbb{1}^T(Bx + b) \leq 1, \forall x, \text{s.t.} Ax \geq a$ is equivalent to

$$1 \geq \max \quad \mathbb{1}^T Bx + \mathbb{1}^T b \tag{B.10}$$
$$\text{s.t.} \quad -Ax \leq -a \tag{B.11}$$

by strong duality, it is equivalent to

$$\begin{aligned} 1 - \mathbb{1}^T b \geq \min \quad & -a^T u \\ \text{s.t.} \quad & -A^T u = B^T \mathbb{1} \\ & u \geq 0 \end{aligned}$$

which reduces to

$$\begin{cases} a^T u \geq -1 + \mathbb{1}^T b \\ A^T u = -B^T \mathbb{1} \\ u \geq 0. \end{cases} \tag{B.12}$$

□

Appendix C

Supplementary material for Chapter 5

C.1 Illustrating Examples on Intertemporal Dependence

In this section, we first discuss three widely used stochastic processes in contextual decision making that are ϕ-mixing. Therefore, these stochastic processes provide examples of circumstances that the previously derived performance guarantees are applicable. Later, we will introduce how the mixing coefficient can be estimated in practice for these examples.

Example C.1.1 (Moving average process)**.** The moving average model of order q, or MA(q) model, is defined to be

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \tag{C.1}$$

where the process ε_t is iid noise with mean 0 and finite variance σ^2, and $\theta_1, \theta_2, \ldots, \theta_q (\theta_q \neq 0)$ are parameters.

It is evident that if an contextual uncertainty process can be characterized as the moving average process, then it is ϕ-mixing. It is because that y_t and y_{t+s} are dependent for any $1 \leq s \leq 1$ and independent for any $s > q$.

Example C.1.2 (Autoregressive process)**.** An autoregressive model of order p, or AR(p) takes the form

$$y_t = \rho_1 t_{t-1} + \rho_2 y_{t-2} + \cdots + \rho_p y_{t-p} + \varepsilon_t \tag{C.2}$$

where the process ε_t is i.i.d noise with mean 0 (not necessarily Gaussian distributed) and finite variance σ^2, and $\rho_1, \rho_2, \ldots \rho_p$ are parameters.

When noises ε_t are bounded, the autoregressive process enjoys the property of ϕ-mixing, under certain conditions [8]. Herein, we state the necessary and sufficient conditions for an autoregressive process to be ϕ-mixing, with bounded innovations.

Theorem C.1.1 (Conditions of AR process to be ϕ-mixing [8]). *The p-th-order autoregressive process defined in C.2 is ϕ-mixing provided:*

(i) $E[\{\log|\varepsilon_1|\}^+] < \infty$,

(ii) the distribution of ε_1 *has a non-trivial absolutely continuous component,*

(iii) d_0 *is independent of* ε_t

(iv) $\{\varepsilon_t\}$ *are i.i.d. random variables*

(v) the root of the characteristic equations $z^p - \rho_1 z^{p-1} - \cdots - \rho_p = 0$ *are less than 1 in modulus*

(vi) exists a finite constant c *such that* $|\varepsilon_1| \leq c$.

On the other hand, [8] shows that an autoregressive process is not ϕ-mixing if the noises have an unbounded support.

Example C.1.3 (Markov modulated process). The Markov modulated process is a widely adopted model that captures a random, fluctuating environment. According to the Markov modulated process, the randomness of $\{Z_t\}_{t=0}^{\infty}$ only depends on the state of an exogenous Markov chain $\{Y_t\}_{t=0}^{\infty}$.

We first review the mixing property of a Markov chain, then show that the ϕ-mixing property of a Markov modulated process is entirely determined by the ϕ-mixing property of the underlying Markov chain, no matter how the distribution of $\{Z_t\}_{t=0}^{\infty}$ depends on Markov chain $\{Y_t\}_{t=0}^{\infty}$. The sufficient and necessary condition for a stationary Markov chain being ϕ-mixing is the uniformly ergodicity (stated in the proof of Theorem 19.1.1 in [63]).

Definition C.1.1 (Uniformly ergodicity). A positive Harris recurrent Markov chain with transition probability kernel P and invariant distribution π is geometrically ergodic if there exists a real number $s < 1$ and a $M < \infty$ on the state space Ω such that

$$\|P^n(x,\cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq M s^n, \qquad x \in \Omega \tag{C.3}$$

where $\|\mu,\nu\|_{\mathrm{TV}} := \sup_{A\in\mathcal{B}} |\mu(A),\nu(A)|$ denotes the total variation distance.

Then Proposition C.1.1.1 demonstrates that the mixing property of the Markov modulated process depends on the mixing property of the exogenous Markov chain.

Proposition C.1.1.1. *The Markov modulated process is ϕ-mixing if the exogenous Markov chain that drives the demand is ϕ-mixing.*

Proof of Proposition C.1.1.1. Let $\sigma_{x,t}$ and $\sigma_{z,t}$ denote the sigma fields of the Markov chain and the contextual uncertainty process, respectively. By considering a larger sigma field, it is clear that

$$\sup_{\substack{n\\A\in\sigma^{\infty}_{z,n+k}\\B\in\sigma^{n}_{z,-\infty}}} \big|\Pr[A|B]-\Pr[A]\big| \le \sup_{\substack{n\\A'\in\sigma(\sigma^{\infty}_{x,n+k},\sigma^{\infty}_{z,n+k})\\B'\in\sigma(\sigma^{\infty}_{x,n+k},\sigma^{n}_{z,-\infty}}} \big|\Pr[A'|B']-\Pr[A']\big|$$

We first consider the events A' and B' that take the form $A'=\tilde{A}\cap A$ and $B'=\tilde{B}\cap B$, respectively. For such events A' and B', we have

$$\sup_{\substack{n\\A'=\tilde{A}\cap A\\\tilde{A}\in\sigma^{\infty}_{x,n+k},A\in\sigma^{\infty}_{d,n+k}\\B'=\tilde{B}\cap B\\\tilde{B}\in\sigma^{\infty}_{x,n+k},B\in\sigma^{n}_{d,-\infty}}} \big|\Pr[A'|B']-\Pr[A']\big| = \sup_{\substack{n\\\tilde{A}\in\sigma^{\infty}_{x,n+k},A\in\sigma^{\infty}_{d,n+k}\\B'=\tilde{B}\cap B\\\tilde{B}\in\sigma^{\infty}_{x,n+k},B\in\sigma^{n}_{d,-\infty}}} \big|\Pr[\tilde{A}\cap A|B']-\Pr[\tilde{A}\cap A]\big| \quad \text{(C.4)}$$

Note that for any n and B', by the definition of $\Pr[\tilde{A}\cap A|B']$ and $\Pr[\tilde{A}\cap A]$, we have the following results

$$\begin{aligned}
&\sup_{\tilde{A}\in\sigma^{\infty}_{x,n+k},A\in\sigma^{\infty}_{d,n+k}} \big|\Pr[\tilde{A}\cap A|B']-\Pr[\tilde{A}\cap A]\big|\\
=&\sup_{\tilde{A}\in\sigma^{\infty}_{x,n+k},A\in\sigma^{\infty}_{d,n+k}} \Big|\int_{\tilde{A}} P(dx|B')P(z\in A|X)-\int_{\tilde{A}}P(dx)P(z\in A|X)\Big|,\\
=&\max\Big\{\sup_{\substack{\tilde{A}\in\sigma^{\infty}_{x,n+k},\\A\in\sigma^{\infty}_{d,n+k}}} \Big|\int_{\tilde{A}}(P(dx|B')-P(dx))P(z\in A|X)\Big|, \sup_{\substack{\tilde{A}\in\sigma^{\infty}_{x,n+k},\\A\in\sigma^{\infty}_{d,n+k}}} \Big|\int_{\tilde{A}}(-P(dx|B')+P(dx))P(z\in A|X)\Big|\Big\}\\
\le&\sup_{\tilde{A}\in\sigma^{\infty}_{x,n+k}} \Big|\int_{\tilde{A}}P(dx|B')-P(dx)\Big| = \sup_{\tilde{A}\in\sigma^{\infty}_{x,n+k}} \big|\Pr[\tilde{A}|B']-\Pr[\tilde{A}]\big|.
\end{aligned}$$

Therefore,

$$(C.4) \le \sup_{\substack{n\\\tilde{A}\in\sigma^{\infty}_{x,n+k}\\\tilde{B}\in\sigma^{\infty}_{x,n+k},B\in\sigma^{n}_{d,-\infty}}} \big|\Pr[\tilde{A}|\tilde{B}\cap B]-\Pr[\tilde{A}]\big| = \sup_{\substack{n\\\tilde{A}\in\sigma^{\infty}_{x,n+k}\\\tilde{B}\in\sigma^{\infty}_{x,n+k}}} \big|\Pr[\tilde{A}|\tilde{B}]-\Pr[\tilde{A}]\big| \quad \text{(C.5)}$$

Then we consider all events A' and B' such that $A'\in\sigma(\sigma^{\infty}_{x,n+k},\sigma^{\infty}_{z,n+k})$ and $B'\in\sigma(\sigma^{\infty}_{x,n+k},\sigma^{n}_{z,-\infty})$. For any $B'\in\sigma(\sigma^{\infty}_{x,n+k},\sigma^{n}_{z,-\infty})$, there are countably many disjoint sets $\tilde{B}_i\cap B_i$ such that $B'=\cup_i\tilde{B}_i\cap B_i$. Moreover, it is possible to make all $\tilde{B}_i$s disjoint. Note that

$$\Pr[A'|B']=\sum_i \frac{\Pr[\tilde{B}_i\cap B_i]}{\Pr[B']}\Pr[A'|\tilde{B}_i\cap B_i],$$

the supreme over $\tilde{B}_i\cap B_i$ must be obtained at a certain $\tilde{B}_i\cap B_i$. Thus,

$$\sup_{\substack{n\\A'\in\sigma(\sigma^{\infty}_{x,n+k},\sigma^{\infty}_{d,n+k})\\B'\in\sigma(\sigma^{\infty}_{x,n+k},\sigma^{n}_{d,-\infty})}} \big|\Pr[A'|B']-\Pr[A']\big| = \sup_{\substack{n\\A'\in\sigma(\sigma^{\infty}_{x,n+k},\sigma^{\infty}_{d,n+k})\\B'=\tilde{B}\cap B\\\tilde{B}\in\sigma^{\infty}_{x,n+k},B\in\sigma^{n}_{d,-\infty}}} \big|\Pr[A'|B']-\Pr[A']\big|.$$

Then for $A' = \cup_i \tilde{A}_i \cap A_i$, where $\tilde{A}_i \cap A_i$s are disjoint and we can make $\tilde{A}_i$s disjoint as well,

$$\sup_{\substack{A' \in \sigma(\sigma^{\infty}_{x,n+k}, \sigma^{\infty}_{d,n+k}) \\ B' = \tilde{B} \cap B \\ \tilde{B} \in \sigma^{\infty}_{x,n+k}, B \in \sigma^{n}_{d,-\infty}}} \big| \Pr[A'|B'] - \Pr[A'] \big| = \sup_{\substack{A' \in \sigma(\sigma^{\infty}_{x,n+k}, \sigma^{\infty}_{d,n+k}) \\ B' = \tilde{B} \cap B \\ \tilde{B} \in \sigma^{\infty}_{x,n+k}, B \in \sigma^{n}_{d,-\infty}}} \big| \sum_i (\Pr[\tilde{A}_i \cap A_i|B'] - \Pr[\tilde{A}_i \cap A_i]) \big|$$

and the supreme can only be achieved either all terms $\Pr[\tilde{A}_i \cap A_i|B'] - \Pr[\tilde{A}_i \cap A_i] \geq 0$ or all terms $\Pr[\tilde{A}_i \cap A_i|B'] - \Pr[\tilde{A}_i \cap A_i] \leq 0$. Therefore,

$$\begin{aligned}
&\sup_{\substack{\tilde{A}_i \cap A_i \\ \tilde{A}_i \in \sigma^{\infty}_{x,n+k}, A \in \sigma^{\infty}_{d,n+k}) \\ B' = \tilde{B} \cap B \\ \tilde{B} \in \sigma^{\infty}_{x,n+k}, B \in \sigma^{n}_{d,-\infty}}} \big| \sum_i (\Pr[\tilde{A}_i \cap A_i|B'] - \Pr[\tilde{A}_i \cap A_i]) \big| \\
= &\sup_{\substack{\tilde{A}_i \cap A_i \\ \tilde{A}_i \in \sigma^{\infty}_{x,n+k}, A \in \sigma^{\infty}_{d,n+k}) \\ B' = \tilde{B} \cap B \\ \tilde{B} \in \sigma^{\infty}_{x,n+k}, B \in \sigma^{n}_{d,-\infty}}} \sum_i \big| \Pr[\tilde{A}_i \cap A_i|B'] - \Pr[\tilde{A}_i \cap A_i] \big| \\
\leq &\sum_i \sup_{\substack{\tilde{A}_i \cap A_i \\ \tilde{A}_i \in \sigma^{\infty}_{x,n+k}, A \in \sigma^{\infty}_{d,n+k}) \\ B' = \tilde{B} \cap B \\ \tilde{B} \in \sigma^{\infty}_{x,n+k}, B \in \sigma^{n}_{d,-\infty}}} \big| \Pr[\tilde{A}_i \cap A_i|B'] - \Pr[\tilde{A}_i \cap A_i] \big| \\
\leq &\sum_i \sup_{\substack{\tilde{A}_i \in \sigma^{\infty}_{x,n+k} \\ \tilde{B} \in \sigma^{\infty}_{x,n+k}}} \big| \Pr[\tilde{A}_i|\tilde{B}] - \Pr[\tilde{A}_i] \big|, \qquad \text{by (C.5)} \\
= &\sup_{\substack{\tilde{A}_i \in \sigma^{\infty}_{x,n+k} \\ \tilde{B} \in \sigma^{\infty}_{x,n+k}}} \big| \Pr[\cup_i \tilde{A}_i|\tilde{B}] - \Pr[\tilde{A}_i] \big| = \sup_{\substack{\tilde{A} \in \sigma^{\infty}_{x,n+k} \\ \tilde{B} \in \sigma^{\infty}_{x,n+k}}} \big| \Pr[\tilde{A}|\tilde{B}] - \Pr[\tilde{A}] \big|.
\end{aligned}$$

The above result states that the ϕ-mixing coefficients of the contextual uncertainty process are less than the ϕ-mixing coefficient of the Markov chain. Hence, if the Markov chain is ϕ-mixing, the contextual process that is modulated based on the Markov chain is also ϕ-mixing. □

Note that evaluating the ϕ-mixing coefficients for both autoregressive and Markov modulated processes reduces to evaluating ϕ-mixing coefficients for Markov chains. Instead of another mixing coefficient, the β-mixing coefficient, to the best of our knowledge, there is no existing results that provide closed-form representations or consistent empirical estimators of $\phi(k)$ for a Markov chain ([100, 101]). In the following, we provide a general recipe for calculating the ϕ-mixing coefficient, $\phi(k)$ of a Markov chain when its transition kernel is given.

Although a closed-form representation is not achievable, [37] provides a closed form upper-bound for $\phi(k)$. According to Proposition 1 of [37], for a stationary homogeneous Markov process X_t with state space $(X, \mathcal{F})$, transition kernel $P_x^t(dy)$, and stationary distribution $\pi(dx)$, there is an upper-bound for ϕ-mixing coefficients:

$$\phi(k) \leq 2 \operatorname*{ess\,sup}_{\pi(dx)} \|P_x^k - \pi\|_{\mathrm{TV}}. \tag{C.6}$$

Recall that it requires uniform ergodicity for a Markov chain to be ϕ-mixing,

$$\phi(k) \leq 2 \operatorname*{ess\,sup}_{\pi(dx)} \|P_x^k - \pi\|_{\mathrm{TV}} \leq M s^k, \tag{C.7}$$

Then the mixing coefficients can be bounded by Ms^k if the uniform ergodic coefficient M is available. On the other hand, when M is not given, the upper-bound of $\phi(k)$ can be directly derived by computing $\|P_x^k - \pi\|_{\mathrm{TV}}$. Particularly, suppose both P_x^k and π are absolutely continuous with respect to another measure μ with the Radon-Nikodym derivatives f_x and f_π respectively, we have $\|P_x^k - \pi\|_{\mathrm{TV}} = \|f_x - f_\pi\|_1$.

Example C.1.4 (Markov chain with finite state space)**.** Consider a Markov chain with state space $\mathcal{S} = \{1, \ldots, N\}, 2 \leq N < \infty$ and transition probability matrix P and stationary distribution π. When P is diagonalizable, [102] provides an approach to calculate M and s:

Proposition C.1.1.2 (Theorem 4.1 of [102])**.** *Suppose β_j, $j = 1, \ldots, N'$ are distinct eigenvalues of P that not equals 1, and $P = W^{-1}\Lambda W$ where Λ is diagonal. Then $s = \max_{j=1,\ldots,M} |\beta_j|$. Furthermore, $M = \max_{i=1,\ldots,N} \sum_{j=1}^{N'} \|e_i \frac{\Pi_{k\neq j}(P-\beta_k I)}{\Pi_{k\neq j}(\beta_j-\beta_k)}\|$, where e_i is the probability measure concentrated at state i.*

C.2 Supplementary Proofs for Section 5.4

Proof of Lemma 5.4.2. Firstly, we bound the probability of $Z_i \neq \tilde{Z}_i$ for each i,

$$\begin{aligned}
\Pr(Z_i \neq \tilde{Z}_i) &= \Pr(\|Z_i\|_\infty > M(n)) \\
&= \Pr(\exp(\|Z_i\|_\infty^a) > \exp M(n)^a) \\
&\leq \frac{A}{\exp(M(n)^a)},
\end{aligned}$$

where the last inequality is a result of Markov inequality and the light-tail property of Z_i. Then $\Pr(S \neq \tilde{S})$ can be bounded by $\frac{nA}{\exp(M^a(n))}$ by applying the union bound. □

Proof of Theorem 5.4.3. Since the upper bound of loss function $\bar{L}(M,n) = b_0 M(n)$, (5.8) becomes

$$\Pr_S\left[\left|R(h_S) - R_{emp}(h_S)\right| > \varepsilon + (6b+2)\beta_0/n + 6b_0 M(n)\phi_0 s^b\right] \leq 2\exp\Big(\frac{-2\varepsilon^2(1-s)^2/(1-s+2\phi_0)^2}{n((b+2)2\frac{\beta_0}{n} + 2b_0\phi_0 M(n)s^b + b_0\frac{M(n)}{n})^2}\Big) + A\exp\big(\ln n - M(n)\big). \quad \text{(C.8)}$$

The first step is to calculate b^*. Since $0 < s < 1$, the term $2b_0\phi_0 M(n)s^b$ decreases as b grows. In the meanwhile, $(b+2)2\frac{\beta_0}{n}$ increases as b increases. By letting the derivative of the two terms equal zero, there is $\frac{b_0}{n} = b_0\phi_0 M(n)\ln\frac{1}{s}s^{b^*}$. By plugging $b = b^*$ to both sides, (C.8) can be represented as

$$\Pr_S\left[\left|R(h_S) - R_{emp}(h_S)\right| > \varepsilon + (6b+2)\beta_0/n + 6\beta_0/(n\ln\frac{1}{s})\right] \leq 2\exp\Big(-\frac{2\varepsilon^2(1-s)^2}{(1-s+2\phi_0)^2}\frac{n}{(\ln(\frac{b_0\phi_0}{\beta_0}nM(n))\frac{2\beta_0}{\ln\frac{1}{s}} + 4\beta_0 + 2\frac{\beta_0}{\ln\frac{1}{s}} + b_0M(n))^2}\Big) + A\exp\big(\ln n - M(n)\big).$$

To balance the convergence rate of the two terms on the right hand side, we choose $M(n) = M_0 n^{1/(a+2)}$ and have

$$\Pr_S\left[\left|R(h_S) - R_{emp}(h_S)\right| > \varepsilon + (6b+2)\beta_0/n + 6\beta_0/(n\ln\frac{1}{s})\right] \leq 2\exp\Big(-\frac{2\varepsilon^2(1-s)^2}{(1-s+2\phi_0)^2}\frac{n}{(\ln(\frac{b_0\phi_0}{\beta_0}M_0 n^{1+1/(a+2)})\frac{2\beta_0}{\ln\frac{1}{s}} + 4\beta_0 + 2\frac{\beta_0}{\ln\frac{1}{s}} + b_0M_0n^{1/(a+2)})^2}\Big) + A\exp\big(\ln n - n^{a/(a+2)}\big).$$

□

C.3 Supplementary Proofs for Section 5.5

In this part, we state the detailed proof of Lemma 5.5.3 and Lemma 5.5.4.

Proof of Lemma 5.5.3. We consider two possible samples S and S^i that only differs in the ith point, and have

$$\Phi(S) - \Phi(S^i) \leq |R_{emp}(h_S) - R_{emp}(h_{S^i})| + |R_{n+r+1}(h_S|S) - R_{n+r+1}(h_{S^i}|S^i)|.$$

First we consider the empirical risk, due to the uniform stability

$$R_{emp}(h_S) - R_{emp}(h_{S^i}) \leq \frac{1}{n}\sum_{j\neq i}|c(h_S,z_j) - c(h_{S^i},z_j)| + \frac{1}{n}|c(h_S,z_i) - c(h_{S^i},z_i')| \leq 2\hat{\beta} + \frac{2M}{n}.$$

Then we consider $\tilde{S}$ being a copy of S that is independent of z_{n+r+1} and S^b being S with the last b points removed. Then we have

$$\begin{aligned}|R_{n+r+1}(h_S|S) - R_{n+r+1}(h_S|\tilde{S})| =& |\mathbb{E}[c(h_{S^b}, z_{n+r+1})|S^b] - \mathbb{E}[c(h_{S^b}, \tilde{z}_{n+r+1})]| + 2b\hat{\beta} \\ \leq& 2b\hat{\beta} + 2M\|P_{n+r+1}(\cdot|\mathcal{F}_{n-b}) - P_{n+r+1}\|_{TV} \\ \leq& 2b\hat{\beta} + M\eta^2_{n-b,n+r+1} \leq M\Pi_{l=n-b+1}^{n+r+1}\theta_l + 2b\hat{\beta}.\end{aligned}$$

The first inequality follows from uniform stability, the second inequality follows from Lemma 2 of [87], and the last inequality is achieved by applying Lemma 5.5.1. Similarly,

$$\begin{aligned}|R_{n+r+1}(h_{S^i}|\tilde{S^i}) - R_{n+r+1}(h_{S^i}|S^i)| =& |\mathbb{E}[c(h_{S^{i,b}}, z_{n+r+1})|S^{i,b}] - \mathbb{E}[c(h_{S^{i,b}}, \tilde{z}_{n+r+1})]| + 2b\hat{\beta} \\ \leq& 2M\|P_{n+r+1}(\cdot|\mathcal{F}_n) - P_{n+r+1}\|_{TV} + 2b\hat{\beta} \\ \leq& M\eta^2_{n-b,n+r+1} + 2b\hat{\beta} \leq M\Pi_{l=n-b+1}^{n+r+1}\theta_l + 2b\hat{\beta}.\end{aligned}$$

Then, by uniform stability,

$$|R_{n+r+1}(h_S|\tilde{S}) - R_{n+r+1}(h_{S^i}|\tilde{S^i})| = |\mathbb{E}[c(h_S, \tilde{z}_{n+r+1})] - \mathbb{E}[c(h_{S^i}, \tilde{z}_{n+r+1})]| \leq 2\hat{\beta}.$$

By combining the above results, we have

$$\begin{aligned}& R_{n+r+1}(h_S|S) - R_{n+r+1}(h_{S^i}|S^i) \\ \leq& |R_{n+r+1}(h_S|S) - R_{n+r+1}(h_S|\tilde{S}) + R_{n+r+1}(h_S|\tilde{S}) - R_{n+r+1}(h_{S^i}|\tilde{S^i}) \\ &\quad + R_{n+r+1}(h_{S^i}|\tilde{S^i}) - R_{n+r+1}(h_{S^i}|S^i)| \\ \leq& |R_{n+r+1}(h_S|S) - R_{n+r+1}(h_S|\tilde{S})| + |R_{n+r+1}(h_{S^i}|\tilde{S^i}) - R_{n+r+1}(h_{S^i}|S^i)| \\ &\quad + |R_{n+r+1}(h_S|\tilde{S}) - R_{n+r+1}(h_{S^i}|\tilde{S^i})| \\ \leq& 2M\Pi_{l=n-b+1}^{n+r+1}\theta_l + (2+4b)\hat{\beta}.\end{aligned}$$

Then the desired results follows:

$$\Phi(S) - \Phi(S^i) \leq (4+4b)\hat{\beta} + 2\frac{M}{n} + 2M\Pi_{l=n-b+1}^{n+r+1}\theta_l.$$

□

Proof of Lemma 5.5.4. For simplicity, we let P_i^j denote the marginal of P on $(\Pi_{k=i}^j \Omega_k, \Pi_{k=i}^j \mathcal{F}_k)$.

$$\mathbb{E}[\Phi(S)] = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}[c(h_S, z_i)] - \mathbb{E}[c(h_S, z_{n+r+1})]) \leq 6b\hat{\beta} + \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}[c(h_{S_b}, z_i)] - \mathbb{E}[c(h_{S_b}, z_{n+r+1})]) \tag{C.9}$$

$$\begin{aligned}
&\leq 6b\hat{\beta} + \frac{6bM}{n} + \frac{1}{n-3b}\sum_{i=b+1}^{n-2b}(\mathbb{E}[c(h_{\tilde{S}_b}, \tilde{z}_i)] - \mathbb{E}[c(h_{\tilde{S}_b}, \tilde{z}_{n+r+1})]) \\
&\quad + \frac{1}{n-3b}\sum_{i=b+1}^{n-2b} 2M\big(\mathbb{E}\|P_i(\cdot|\mathcal{F}_{i-b}) - P_i\|_{TV} + \mathbb{E}\|P_{i+b}^{n-b}(\cdot|\mathcal{F}_i) - P_{i+b}^{n-b}\|_{TV} \\
&\qquad\qquad + \mathbb{E}\|P_{r+n+1}(\cdot|\mathcal{F}_{n-b}) - P_{r+n+1}\|_{TV}\big)
\end{aligned} \tag{C.10}$$

$$\begin{aligned}
&\leq 6b\hat{\beta} + \frac{6bM}{n} + \frac{2M}{n-3b}\sum_{i=b+1}^{n-2b}(\Pi_{l=i-b+1}^{i}\theta_l + \Pi_{l=i+1}^{i+b}\theta_l + \Pi_{l=n-b+1}^{n+r+1}\theta_l) \\
&\quad + \frac{1}{n-3b}\sum_{i=b}^{n-2b}(\mathbb{E}[c(h_{\tilde{S}_b}, \tilde{z}_i)] - \mathbb{E}[c(h_{\tilde{S}_b}, \tilde{z}_{n+r+1})])
\end{aligned} \tag{C.11}$$

$$\leq 6b\hat{\beta} + \frac{2M}{n-3b}\sum_{i=b+1}^{n-2b}(\Pi_{l=i-b+1}^{i}\theta_l + \Pi_{l=i+1}^{i+b}\theta_l + \Pi_{l=n-b+1}^{n+r+1}\theta_l) + \frac{1}{n-3b}\sum_{i=b}^{n-2b}\|P_{n+r+1} - P_i\|_{TV} \tag{C.12}$$

$$\leq 6b\hat{\beta} + \frac{2M}{n-3b}\sum_{i=b+1}^{n-2b}(\Pi_{l=i-b+1}^{i}\theta_l + \Pi_{l=i+1}^{i+b}\theta_l + \Pi_{l=n+1}^{n+r+1}\theta_l) + \Delta$$

Inequality (C.9) follows from the uniform stability and (C.10) is achieved by considering four independent blocks split by three gaps with length b and apply Lemma 2 of [87]. Inequality (C.11) and (C.12) follow from Lemma 5.5.1. □